B

预测信贷违约的探究

——基于德国信贷数据集

2018110760 李晨茜

一、研究背景与问题

1. 研究背景

信贷是货币持有者将约定数额的资金按约定的利率暂时借出，借款者在约定期限内，按约定的条件还本付息的信用活动。

在现实生活中有些人由于没有或者较少的信用记录，很难获得贷款，为了增加对与没有银行帐号或信用记录人群的借贷的包容性，信贷机构会利用各种替代数据：电信或交易信息等等客户的历史的行为数据来预测客户还款能力 .基于这些数据，利用各种方法来做出这些预测，确保有能力还款的客户不会被拒绝、并且避免与违约风险高的客户进行交易从而使得银行或信贷机构利润最大化。

2. 研究问题

本报告希望研究贷款者是否违约的影响因素及影响程度，并通过建立模型尽可能根据已知信息预测贷款者的违约与否。

二、数据预处理

一开始先要介绍这个数据集有多少条观测，多少个变量

首先对数据集中涉及的变量进行整理。

因为自变量中有多个变量存在大于两个取值，为避免后续回归模型中的虚拟变量陷阱，分别为它们设置基准组。设置 A11 为支票账户状态（checkingstatus1）的基准组；设置 A30 为信贷历史（history）的基准组；设置 A410 为贷款目的（purpose）的基准组；设置 A61 为存款数（savings）的基准组；设置 A71 为工作情况（employ）的基准组；设置 1 为分期付款率占可支配收入的百分比(installment)的基准组;设置 A91 为个人婚姻状况和性别(status)的基准组；设置 A101 为其他担保人（others）的基准组；设置 1 为居住年数（residence）的基准组；设置 A124 为财产情况（property）的基准组；设置 A143 为其他还款计划（otherplans）的基准组；设置 A151 为住房情况（housing）的基准组；设置 A171 为工作情况（job）的基准组;

表格 1 变量说明

变量名		中文解释	类别	备注	基准组
因变量	default	是否违约	分类变量	1=是，0=否	
自变量	checkingstatus1	支票账户状态	分类变量	A11：<0 DM, A12：0 <= x <200 DM, A13：>= 200 DM /至少一年的薪水分配，A14：无支票帐户	A11
	duration	持续时间	连续型变量	月	
	history	信贷历	分类	A30：未提取任何信用/已全额偿还所有信	A30

	史	变量	用额，A31：已偿还该银行的所有信用额，A32：已到期已偿还的现有信用额，A33：过去的还款延迟，A34：关键帐户/其他信用额现有（不在此银行）	
purpose	借款目的	分类变量	A40：新车；A41：二手车；A42：家具/设备；A43：收音机/电视；A44：家用电器 A45：维修；A46：教育；A47：度假；A48：再培训、再教育；A49：商业；A410：其他	A410
amount	借款金额	连续型变量		
savings	存款数	分类变量	A61：<100 DM，A62：100 <= x <500 DM，A63：500 <= x <1000 DM，A64：> = 1000 DM，A65：未知/无储蓄账户	A61
employ	工作年数情况	分类变量	A71：待业，A72：<1 年，A73：1 <= x <4 年，A74：4 <= x <7 年，A75：..> = 7 年	A71
installment	分期付款率占可支配收入的百分比	分类变量	取值 1,2,3,4	1
status	个人婚姻状况和性别	分类变量	A91：男性：离婚/分居，A92：女性：离婚/分居/已婚，A93：男性：单身，A94：男性：已婚/丧偶，A95：女性：单身	A91
others	其他担保人	分类变量	A101：无，A102：共同申请人，A103：担保人	A101
residence	至今居住	分类变量	居住年数	1
property	财产情况	分类变量	A121：不动产，A122：如果不是 A121，那么建筑协会储蓄协议/人寿保险，A123：如果不是 A121 / A122，不是属性 6 的汽车或其他；A124：未知/没有财产	A124
age	年龄	连续型变量		
otherplans	其他分期付款计划	分类变量	A141：银行，A142：商店，A143：无	A143
housing	住房情况	分类变量	A151:租房，A152:自有，A153:免费	A151
cards	该银行现有信贷的数量	分类变量		

job	工作情况	分类变量	A171:失业/非技术人员/非居 A172:非技术人员-居民 A173:技术人员/官员 A174:管理/个体经营/高级的员工/官员	A171
liable	还款人数	分类变量		
tele	是否有电话	分类变量	A191:无，A192:有，登记在客户名下	
foreign	是否外国劳工	分类变量	A201: 有, A202: 无	

经检查，该数据集中不存在缺失值和异常值。

三、数据可视化

为了初步观察探究不同自变量对因变量的影响，画出分类自变量和因变量的条形图和连续自变量和因变量之间的分类箱线图如下：

图 1 分类自变量与因变量可视化 1

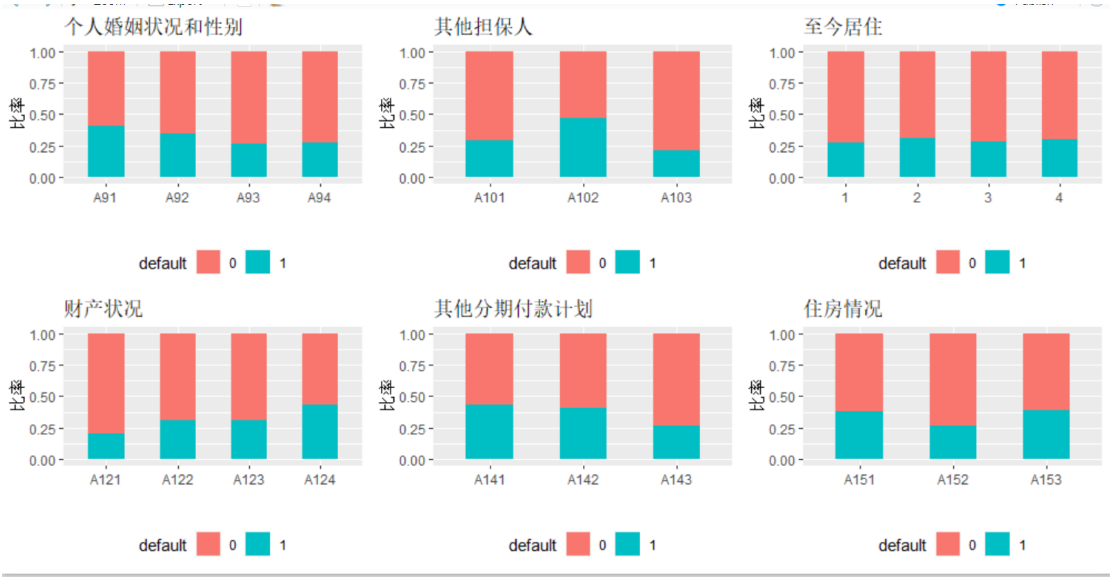

图 2 分类自变量与因变量可视化 2

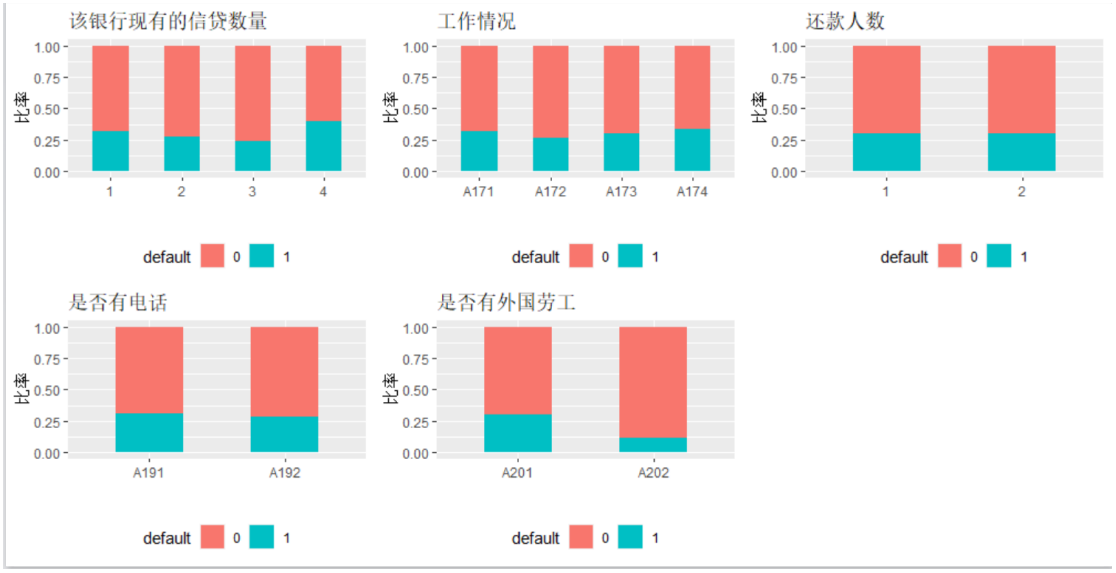

图 3 分类自变量与因变量可视化 3

通过绘制条形图可以发现:

随着支票账户金额的增加，违约概率有降低的趋势，符合常理；而没有支票账户的客户违约率最低，考虑可能是由于没有支票账户的客户有更小的消费动力，更有可能有足够现金还款。

从存款数来看，违约率有随着存款数上升而下降的趋势，说明存款越多的人违约的概率就越小。

从工作年数来看，工作小于一年的和无业人员的违约率都很高，结合常识分析，无业人员由于没有稳定的收入，极可能出现入不敷出的情况，导致违约；而猜测工作小于一年的人可能由于初入职场，对收入没有很好的规划，导致出现信贷问题。其余三个选项显示随着工作年数的增加违约率是在降低的，符合常情常理。

从个人婚姻状况和性别来看，是否违约在同婚姻条件不同性别的客户群体中差别不大，而总的来说离婚/分居的客户群体相较于已婚/单身的客户群体违约率更高，猜测可能婚姻离异给他们的生活和财务状况带来了一定的冲击(或者由于他们个人的状况比较糟糕导致的离婚)，从而致使违约率提高。

从担保人情况来看，共同贷款者情况的客户违约概率最高，考虑有可能是因为两个申请者之间可能存在互相推卸责任的情况，导致违约。

从财产状况来看，拥有不动产的客户违约率最低，而无资产的客户违约率最高，其中随着客户资产额的减少，违约率有提高的趋势。

从其他分期付款计划来看，有其他银行或商店贷款的客户违约率高于没有其他分期付款计划的客户。猜测可能是由于贷款条目比较多的客户容易出现资金链断裂的问题，且过多的贷款项目也在某种程度上反映了该客户不佳的经济状况。

从住房情况来看，拥有自己住房的客户违约率最低，和前文财产状况反映的趋势一致，拥有不动产的客户的资产状况更可能好，违约的可能性更低。

从国籍来看，外国劳工的违约率相较于非外国劳工更高。猜测可能存在两方面原因，一方面，外国劳工在本国的就业难度要高于本国居民，导致他们更难有稳定的收入；另一方面，外国劳工可能不在本国的信用系统内，缺少对于其违约的约束和守约的激励。

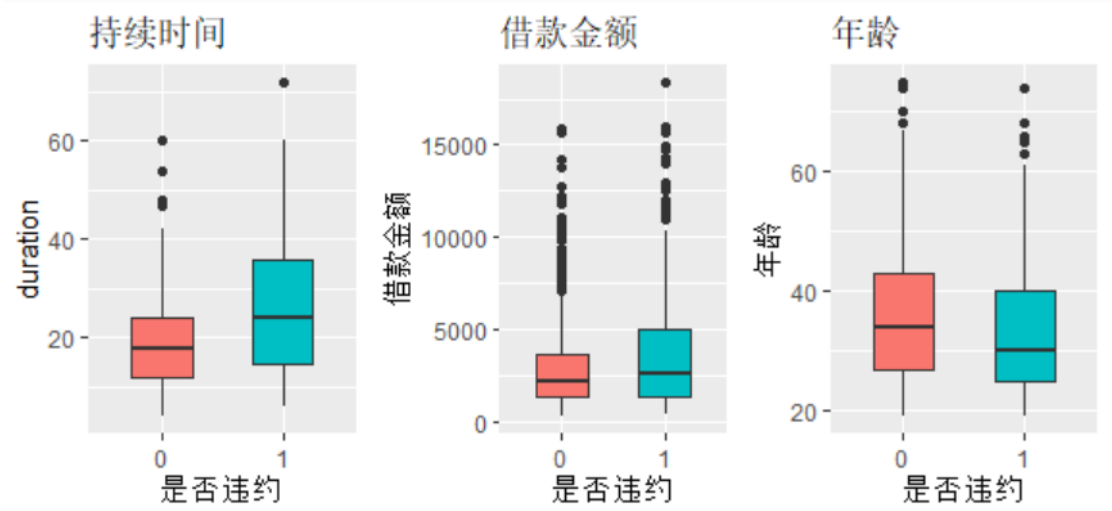

图 4 连续自变量与因变量可视化

从分类箱线图可以看出持续时间越长的贷款越容易被贷款人违约，可能由于更长的时间给银行带来了更多的不确定性和风险。金额越高的贷款越容易遭到违约，可能由于金额较大导致贷款人难以在规定还款时间内集齐足够的现金用来还款。而年龄较低的还款人更容易违约，可能由于年龄较低的客户生活、工作都还没有进入稳定的状态，自身的财产管理能力相较于年长的客户也较差一些，更容易出现财务问题，导致违约。

四、logistic 模型

1. 模型说明

由于本数据集的因变量（是否违约）是分类变量，如果仍采用普通的线性回归不仅误差可能不服从正态分布，而且预测值往往是连续的，并极有可能超出[0,1]的范围。

考虑到以上问题，对本数据采用 Logistic 回归模型：

如果针对Y而言，是指预测值往往不为0或1

$$\text{logit}(p(\boldsymbol{x})) = \log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \boldsymbol{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

其中，模型的回归参数采用极大似然估计，对数似然函数可以写成：

$$l(\beta) = \sum_{i=1}^{m}\left[y_i \log\left(\frac{p_i}{1-p_i}\right) + n_i \log(1-p_i)\right]$$

通过该模型，可以根据已有自变量预测出发生违约的概率有多大。

2. 模型建立

以是否违约作为因变量，从原有数据集的 1000 条观测中抽取 900 条作为训练集拟合模型，得到结果如下表所示（模型 1）：

表格 2 模型 1 结果

变量名	系数	P 值	备注
(INTERCEPT)	-1.049	0.455435	
CHECKINGSTATUS1A12	-0.4659	0.047047	
CHECKINGSTATUS1A13	-1.082	0.007202	
CHECKINGSTATUS1A14	-1.909	<0.001	

DURATION	0.03389	<0.001
HISTORYA31	0.2246	0.710334
HISTORYA32	-0.442	0.365643
HISTORYA33	-0.9944	0.055352
HISTORYA34	-1.499	0.002014
PURPOSEA40	1.591	0.060013
PURPOSEA41	-0.1536	0.861523
PURPOSEA42	0.7039	0.409796
PURPOSEA43	0.6711	0.430472
PURPOSEA44	1.155	0.306037
PURPOSEA45	1.011	0.319988
PURPOSEA46	1.845	0.043381
PURPOSEA48	-0.3091	0.832077
PURPOSEA49	0.9849	0.262104
AMOUNT	9.775E-05	0.050665
SAVINGSA62	-0.4974	0.113038
SAVINGSA63	-0.0217	0.957598
SAVINGSA64	-1.521	0.01026
SAVINGSA65	-0.9327	<0.001
EMPLOYA72	0.104	0.825897
EMPLOYA73	-0.2039	0.649265
EMPLOYA74	-0.7086	0.151807
EMPLOYA75	-0.2086	0.643051
INSTALLMENT2	0.1679	0.604423
INSTALLMENT3	0.5217	0.145337
INSTALLMENT4	0.8617	0.006867
STATUSA92	-0.2032	0.618719
STATUSA93	-0.895	0.025825
STATUSA94	-0.4186	0.387838
OTHERSA102	0.4244	0.341311
OTHERSA103	-0.9037	0.038649
RESIDENCE2	0.6671	0.036187
RESIDENCE3	0.362	0.31514
RESIDENCE4	0.3981	0.221405
PROPERTYA121	-0.7011	0.132627
PROPERTYA122	-0.4363	0.339206
PROPERTYA123	-0.4614	0.29837
AGE	-0.01946	0.05626
OTHERPLANSA142	0.5913	0.160281
OTHERPLANSA141	0.5427	0.033726
HOUSINGA152	-0.4259	0.091782
HOUSINGA153	-0.4832	0.351708
CARDS2	0.6122	0.021882
CARDS3	0.4909	0.436068

尽量保留四位小数保持一致。

CARDS4	1.363	0.257039
JOBA172	1.054	0.173025
JOBA173	0.943	0.210456
JOBA174	0.9988	0.19547
LIABLE2	0.3848	0.161029
TELEA192	-0.1913	0.380758
FOREIGNA202	-1.417	0.037897

可以发现在 5%的显著性水平下有很多变量都不显著，利用 AIC 逐步回归对模型中包含的变量进行选择和剔除，结果如下（模型 2）：

表格 3 模型 2 结果

变量	系数	EXP（系数）	P 值	基准组
(INTERCEPT)	-0.3597	0.6978857	0.741031	
CHECKINGSTATUS1A12	-0.479	0.6194025	0.03306	A11
CHECKINGSTATUS1A13	-1.135	0.3214221	0.003776	
CHECKINGSTATUS1A14	-1.867	0.1545867	1.59E-14	
DURATION	0.03133	1.031826	0.001037	
HISTORYA31	-0.1972	0.8210264	0.723951	A30
HISTORYA32	-0.8415	0.4310634	0.058317	
HISTORYA33	-1.02	0.3605949	0.041111	
HISTORYA34	-1.562	0.2097162	0.000847	
PURPOSEA40	1.599	4.9480819	0.044768	A410
PURPOSEA41	0.00462	1.0046307	0.995585	
PURPOSEA42	0.7574	2.1327239	0.344736	
PURPOSEA43	0.6229	1.8643268	0.436242	
PURPOSEA44	1.034	2.8122925	0.337385	
PURPOSEA45	1.029	2.7982662	0.28411	
PURPOSEA46	1.951	7.0357198	0.023909	
PURPOSEA48	-0.3588	0.698514	0.804277	
PURPOSEA49	0.868	2.3821418	0.290823	
AMOUNT	0.0001102	1.0001102	0.01501	
SAVINGSA62	-0.4399	0.6441008	0.133283	A61
SAVINGSA63	-0.1619	0.8505263	0.685202	
SAVINGSA64	-1.488	0.2258239	0.008239	
SAVINGSA65	-0.9673	0.3801079	0.000434	
INSTALLMENT2	0.1643	1.1785678	0.600314	1
INSTALLMENT3	0.5601	1.7508476	0.10504	
INSTALLMENT4	0.9121	2.4895451	0.002646	
STATUSA92	-0.13	0.8780954	0.738318	A91
STATUSA93	-0.8609	0.4227814	0.025754	
STATUSA94	-0.4539	0.6351462	0.331371	
OTHERSA102	0.5431	1.7213347	0.199304	A101
OTHERSA103	-0.985	0.3734392	0.019727	

AGE	-0.01895	0.9812284	0.029965	
OTHERPLANSA142	0.5472	1.7284067	0.180667	A143
OTHERPLANSA141	0.5841	1.7933762	0.018194	
LIABLE2	0.4239	1.5279088	0.107277	
FOREIGNA202	-1.233	0.291417	0.065297	

这里关于系数的解读都不对

该模型说明：

控制其他因素不变，支票账户状态大于 0 小于 200DM 的客户的违约概率是支票账户状态小于 0 的客户的 0.62 倍，支票账户状态大于 200DM 或大于一年的薪水分配的客户的违约概率是支票账户状态小于 0 的客户的 0.32 倍，无支票账户的客户的违约概率是支票账户状态小于 0 的客户的 0.15 倍；

控制其他因素不变，贷款持续时间每增加一个月，违约概率是原来的 1.03 倍；

控制其他因素不变，已偿还该银行所有贷款的客户的违约概率是已全额偿还所有贷款/未贷款的客户的 0.82 倍，至今按期还款的客户的违约概率是已全额偿还所有贷款/未贷款的客户的 0.43 倍，过去曾延迟还款的客户的违约概率是已全额偿还所有贷款/未贷款的客户的 0.36 倍，有其他未还贷款的客户的违约概率是已全额偿还所有贷款/未贷款的客户的 0.21 倍；

控制其他因素不变，贷款目的为买新车的客户的违约概率是目的为其他的客户的 4.95 倍，贷款目的为买二手车的客户的违约概率是目的为其他的客户的 1.004 倍，贷款目的为买家具/设备的客户的违约概率是目的为其他的客户的 2.13 倍，贷款目的为买收音机/电视机的客户的违约概率是目的为其他的客户的 1.86 倍，贷款目的为买家用电器的客户的违约概率是目的为其他的客户的 2.81 倍，贷款目的为修复的客户的违约概率是目的为其他的客户的 2.8 倍，贷款目的为教育的客户的违约概率是目的为其他的客户的 7.04 倍，贷款目的为支付再教育再培训的客户的违约概率是目的为其他的客户的 0.7 倍，贷款目的为商业用途的客户的违约概率是目的为其他的客户的 2.38 倍。

控制其他因素不变，贷款金额每增加 1，违约概率变为原来的 1.00011 倍。

控制其他因素不变，储蓄账户金额在 100DM 和 500DM 之间的客户的违约概率是储蓄账户金额小于 100DM 的 0.64 倍，储蓄账户金额在 500DM 和 1000DM 之间的客户的违约概率是储蓄账户金额小于 100DM 的 0.85 倍，储蓄账户金额大于 1000DM 的客户的违约概率是储蓄账户金额小于 100DM 的 0.23 倍，没有储蓄账户的客户的违约概率是储蓄账户金额小于 100DM 的 0.38 倍。

控制其他因素不变，分期付款率占可支配收入的百分比为 2 的客户的违约概率为分期付款率占可支配收入的百分比为 1 的 1.17 倍，分期付款率占可支配收入的百分比为 3 的客户的违约概率为分期付款率占可支配收入的百分比为 1 的 1.75 倍，分期付款率占可支配收入的百分比为 3 的客户的违约概率为分期付款率占可支配收入的百分比为 1 的 2.49 倍。

控制其他因素不变，离异/分居女性客户的违约概率是离异/分居男性客户的 0.88 倍，单身男性客户的违约概率是离异/分居男性客户的 0.42 倍，已婚/丧偶男性客户的违约概率是离异/分居男性客户的 0.64 倍，单身女性客户的违约概率是离异/分居男性客户的 0.64 倍。

控制其他因素不变，客户年龄每上涨 1，违约概率变为原来的 0.98 倍。

控制其他因素不变，有商店贷款的客户的违约概率是没有其他贷款的客户的 1.73 倍，有银行贷款的客户的违约概率是没有其他贷款的客户的 0.37 倍。

控制其他因素不变，两人共同还款的客户的违约概率是一人还款的 1.53 倍。

控制其他因素不变，非外国劳工客户的违约概率是外国劳工的 0.29 倍。

对该模型进行卡方检验，通过比较拟合值和真实值差距来判断该模型的拟合优度。得到 p 值约等于 0.8898，在 5%的显著性水平下不拒绝原假设，认为拟合值和真实值没有显著差

是Deviance检验吧

别。

为了探究每个模型中每个变量的显著性，对每个变量做卡方检验，得到结果如下表所示：

表格 4 模型 2 变量系数的卡方检验

变量名	PR(>CHI)
CHECKINGSTATUS1	< 0.001
DURATION	<0.001
HISTORY	<0.001
PURPOSE	<0.001
AMOUNT	0.286573
SAVINGS	0.001213
INSTALLMENT	0.018217
STATUS	0.004151
OTHERS	0.01656
AGE	0.055188
OTHERPLANS	0.049166
LIABLE	0.116943
FOREIGN	0.040259

发现在 5%的显著性水平下，只有变量 amount，age，liable 不显著，而这三个变量中，最不显著的是 liable。为了使模型变得更有效，尝试去掉不显著的变量。去掉变量 liable 之后得到的模型如下表所示（模型 3）：

表格 5 模型 3 结果

变量	参数估计	P 值
(INTERCEPT)	-0.32341	0.765772
CHECKINGSTATUS1A12	-0.49874	0.026222
CHECKINGSTATUS1A13	-1.16099	0.003105
CHECKINGSTATUS1A14	-1.86991	1.27E-14
DURATION	0.030731	0.001246
HISTORYA31	-0.17002	0.760442
HISTORYA32	-0.85247	0.053784
HISTORYA33	-0.99139	0.046313
HISTORYA34	-1.5694	0.000761
PURPOSEA40	1.609382	0.043217
PURPOSEA41	0.039586	0.962123
PURPOSEA42	0.746633	0.351203
PURPOSEA43	0.614333	0.44207
PURPOSEA44	0.991592	0.356374
PURPOSEA45	1.03219	0.284039
PURPOSEA46	1.972655	0.02237
PURPOSEA48	-0.30763	0.832708
PURPOSEA49	0.864685	0.292095
AMOUNT	0.000108	0.017103
SAVINGSA62	-0.44205	0.130215
SAVINGSA63	-0.16844	0.672624

SAVINGSA64	-1.41353	0.010665
SAVINGSA65	-0.96101	0.000465
INSTALLMENT2	0.152417	0.627648
INSTALLMENT3	0.531261	0.124035
INSTALLMENT4	0.863591	0.004297
STATUSA92	-0.11987	0.75833
STATUSA93	-0.75662	0.046771
STATUSA94	-0.45381	0.332058
OTHERSA102	0.504548	0.232976
OTHERSA103	-0.95779	0.023476
AGE	-0.01791	0.037972
OTHERPLANSA142	0.51141	0.210106
OTHERPLANSA141	0.578579	0.019379
FOREIGNA202	-1.22916	0.067948

p>0.05是拒绝原假设?

对该更新的模型进行嵌套模型的似然比检验，得到 p 值为 0.109676>0.05，说明在 5%的显著性水平下，拒绝原假设，认为原模型不是冗余的。并且对模型 3 进行卡方检验之后发现 p 值为 0.8821。而对模型 2（逐步回归后的）进行卡方检验的 p 值为 0.8898。且模型 3 的 AIC 为 885.97，模型 2 的 AIC 为 885.41。这些都说明对模型 2 删除变量是没有必要的，虽然模型 2 中有不显著的变量，但这些变量并不是冗余的。因此模型 2 是合理的。

3. 预测——基于测试集

测试集为原数据集中除用来拟合模型的 900 个观测之外的 100 个观测，用该测试集的自变量作为已知信息代入模型对是否违约进行预测。可以用 ROC 曲线来说明预测效果，ROC 曲线（经平滑处理后）如下图所示：

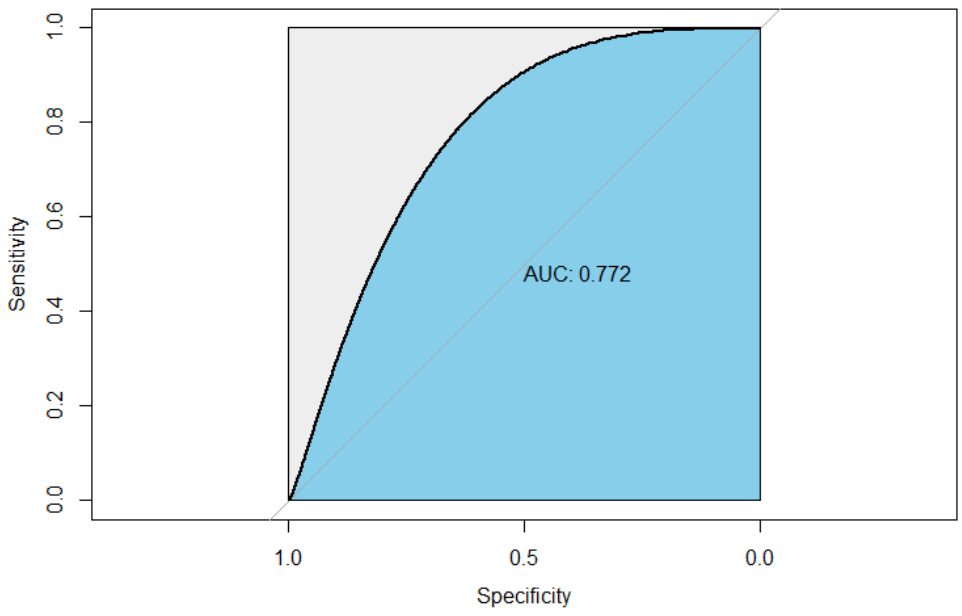

图 5 ROC 结果

由上图可知 AUC=0.772>0.5，说明本模型优于随机猜测，只要妥善设定阈值，就可以有

预测价值。 没有总结吗?

五、附录代码

```
library(gridExtra)
library(ggplot2)
library(mice)
library(dplyr)
library(pROC)

credit_origin<-read.csv("D://学习//大三上//探索性数据分析
//hw2//germancredit.csv",header=T)
id<-c(1:1000)
credit_origin<-data.frame(id,credit_origin)
credit<-sample_n(credit_origin,900,replace = FALSE)
credittest =
subset(credit_origin, !credit_origin$id%in%c(credit$id)) #从一个数据
框中去掉另一个数据框
credit<-credit[,-1]
credittest<-credittest[,-1]
#设定基准组
credit$checkingstatus1<-
factor(as.character(credit$checkingstatus1),levels=c("A11","A12","A
13","A14"))
credit$history<-
factor(as.character(credit$history),levels=c('A30','A31','A32','A33
','A34'))
credit$purpose<-
factor(as.character(credit$purpose),levels=c('A410','A40','A41','A4
2','A43','A44','A45','A46','A47','A48','A49'))
credit$savings<-
factor(as.character(credit$savings),levels=c('A61','A62','A63','A64
','A65'))
credit$employ<-
factor(as.character(credit$employ),levels=c('A71','A72','A73','A74'
,'A75'))
credit$installment<-
factor(as.character(credit$installment),levels=c('1','2','3','4'))
credit$status<-
factor(as.character(credit$status),levels=c('A91','A92','A93','A94'
,'A95'))
credit$others<-
factor(as.character(credit$others),levels=c('A101','A102','A103'))
credit$residence<-
factor(as.character(credit$residence),levels=c('1','2','3','4'))
```

```
credit$property<-
factor(as.character(credit$property),levels=c('A124','A121','A122',
'A123'))
credit$otherplans<-
factor(as.character(credit$otherplans),levels=c('A143','A142','A141
'))
credit$housing<-
factor(as.character(credit$housing),levels=c('A151','A152','A153'))
credit$job<-
factor(as.character(credit$job),levels=c('A171','A172','A173','A174
'))
#把数据集中的分类自变量都 factor
credit_origin$default<-as.factor(credit_origin$default)
credit_origin$checkingstatus1<-
as.factor(credit_origin$checkingstatus1)
credit_origin$history<-as.factor(credit_origin$history)
credit_origin$purpose<-as.factor(credit_origin$purpose)
credit_origin$savings<-as.factor(credit_origin$savings)
credit_origin$employ<-as.factor(credit_origin$employ)
credit_origin$installment<-as.factor(credit_origin$installment)
credit_origin$status<-as.factor(credit_origin$status)
credit_origin$others<-as.factor(credit_origin$others)
credit_origin$residence<-as.factor(credit_origin$residence)
credit_origin$property<-as.factor(credit_origin$property)
credit_origin$otherplans<-as.factor(credit_origin$otherplans)
credit_origin$housing<-as.factor(credit_origin$housing)
credit_origin$cards<-as.factor(credit_origin$cards)
credit_origin$job<-as.factor(credit_origin$job)
credit_origin$liable<-as.factor(credit_origin$liable)
credit_origin$tele<-as.factor(credit_origin$tele)
credit_origin$foreign<-as.factor(credit_origin$foreign)
str(credit_origin)
md.pattern(credit_origin)

#分类自变量可视化
p1<-
ggplot(credit,aes(x=checkingstatus1,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='支票账户状况')+theme(legend.position =
"bottom")
p2<-ggplot(credit,aes(x=history,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='信贷历史')+theme(legend.position =
"bottom")
```

```
p3<-ggplot(credit,aes(x=purpose,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='借款目的')+theme(legend.position =
"bottom")
p4<-ggplot(credit,aes(x=savings,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='存款数')+theme(legend.position =
"bottom")
p5<-ggplot(credit,aes(x=employ,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='工作年数情况')+theme(legend.position =
"bottom")
p6<-ggplot(credit,aes(x=installment,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='分期付款率占可支配收入的百分比
')+theme(legend.position = "bottom")
p7<-ggplot(credit,aes(x=status,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='个人婚姻状况和性别')+theme(legend.position
= "bottom")
p8<-ggplot(credit,aes(x=others,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='其他担保人')+theme(legend.position =
"bottom")
p9<-ggplot(credit,aes(x=residence,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='至今居住')+theme(legend.position =
"bottom")
p10<-ggplot(credit,aes(x=property,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='财产状况')+theme(legend.position =
"bottom")
p11<-ggplot(credit,aes(x=otherplans,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='其他分期付款计划')+theme(legend.position =
"bottom")
p12<-ggplot(credit,aes(x=housing,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='住房情况')+theme(legend.position =
"bottom")
p13<-ggplot(credit,aes(x=cards,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='该银行现有的信贷数量
')+theme(legend.position = "bottom")
```

```
p14<-ggplot(credit,aes(x=job,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='工作情况')+theme(legend.position =
"bottom")
p15<-ggplot(credit,aes(x=liable,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='还款人数')+theme(legend.position =
"bottom")
p16<-ggplot(credit,aes(x=tele,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='是否有电话')+theme(legend.position =
"bottom")
p17<-ggplot(credit,aes(x=foreign,fill=default))+geom_bar(width =
0.5,position='fill')+
  labs(x="",y="比率",title='是否有外国劳工')+theme(legend.position =
"bottom")
grid.arrange(p1,p2,p3,p4,p5,p6,ncol=3,newpage = T)
grid.arrange(p7,p8,p9,p10,p11,p12,ncol=3,newpage = T)
grid.arrange(p13,p14,p15,p16,p17,ncol=3,newpage = T)

#连续自变量可视化
p18<-
ggplot(credit,aes(x=default,y=duration,fill=default))+geom_boxplot(
width = 0.5,show.legend = FALSE)+
  labs(x="是否违约",y="duration",title='持续时间
')+theme(legend.position = "bottom")
p19<-
ggplot(credit,aes(x=default,y=amount,fill=default))+geom_boxplot(wi
dth = 0.5,show.legend = FALSE)+
  labs(x="是否违约",y="借款金额",title='借款金额
')+theme(legend.position = "bottom")
p20<-
ggplot(credit,aes(x=default,y=age,fill=default))+geom_boxplot(width
= 0.5,show.legend = FALSE)+
  labs(x="是否违约",y="年龄",title='年龄')+theme(legend.position =
"bottom")
grid.arrange(p18,p19,p20,ncol=3,newpage = T)

#全模型放入 AIC 的 logistic 回归
#model
glm1<-glm(default~.,family=binomial(link="logit"),data=credit)
glm1_steped<-step(glm1,k=2)
summary(glm1)
summary(glm1_steped)
```

```r
#卡方检验
pchisq(glm1_steped$deviance,glm1_steped$df.residual,lower.tail = FALSE)
anova(glm1_steped,test='Chisq') #发现 amount,age,others,housing 在 0.05 的显著性水平下不显著

#推断
exp(glm1_steped$coefficients) #OR 估计

#预测
pre<-predict(glm1_steped,credittest,type="response")

#预测效果（ROC）
roc<-roc(credittest$default,pre)
plot(smooth(roc),print.auc=TRUE,auc.polygon=TRUE,grid.col=c("green","red"),max.auc.polygon=TRUE,auc.polygon.col="skyblue")


#去掉 amount,age,liable
glm2<-update(glm1_steped,.~.-liable)
summary(glm2)
#卡方检验
pchisq(glm2$deviance,glm2$df.residual,lower.tail = FALSE)
#似然比检验（似然比检验的零假设：固定效应模型是冗余的）
lrt<--2*(logLik(glm2)[1]-logLik(glm1_steped)[1])
pchisq(lrt,df=1,lower.tail = FALSE)#拒绝原假设
```