信用卡评分建模分析

2020年5月

摘 要

本文主要对信用卡进行评分建模分析。首先，我们对数据进行预处理，删去异常值，并且对于少量缺失的数据用众数填充，大量缺失的数据用随机森林填充。相关性分析方面，我们分别采用定性分析和定量分析。定性分析中，关系明显的用箱线图呈现，箱线图不明显的用违约概率的折线图来进一步展示。定量分析中，在分析是否违约和其它特征的相关性时，采用点二列相关系数；在分析其它特征之间的相关性时，采用皮尔逊相关系数。不同类型的变量采取不同的指标评估，让结果更加科学合理。因变量分析中，我们发现样本十分不均衡，这样会导致预测效果极差，因此我们分别采用过采样与降采样的方法进行调整。模型选择上，我们采用适合分类的逻辑斯蒂回归模型和并行实现简单、不易过拟合的随机森林模型，并应用于过采样的数据中，结合变量进行优化，最后均达到了很好的效果：准确率约78%，违约样本中准确率在80%以上，违约风险在2%以下。最后，我们将这两个模型用于降采样的数据中，发现逻辑斯蒂回归模型效果很差，随机森林模型尚可。

关键词：信用卡违约、逻辑斯蒂回归模型、随机森林模型、过采样与降采样、相关性

一、引言

1.1 问题描述

信用评分技术是一种常见的应用统计模型，是一种对贷款申请人（信用卡申请人）做风险评估分值的方法。有了信用评分技术，在借贷的过程中，银行就可以对风险进行有效的评估，从而最大程度地规避信用风险。

信用风险这个概念是受信人不能履行还本付息的责任而使授信人的预期收益与实际收益发生偏离的可能性的度量标准。信用风险是在交易的过程中交易对手未能履行约定合同中的义务造成经济损失的风险。借贷场景中的评分卡是一种以分数的形式来衡量风险几率的一种手段，也是对未来一段时间内违约、逾期、失联概率的预测。本项目对评分卡中的分数的定义是，在风险越小的情况下，得到的评分越高。

1.2 数据来源

Kaggle中的Give Me Some Credit的数据。

1.3 问题限定

本组项目讨论的是申请者评级模型，即主要应用于相关融资类业务中新用户的主体评级。根据申请者的历史数据，对客户的信用进行评估，以决定是否同意其贷款申请。

二、数据预处理

对一个未知的数据集，可能存在数据缺失、存在异常值、数据重复等问题，需要在建立模型前对数据进行处理。

2.1 缺失值处理

首先得到数据的整体分布：

```
#   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
0   Unnamed: 0                            150000 non-null  int64
1   SeriousDlqin2yrs                      150000 non-null  int64
2   RevolvingUtilizationOfUnsecuredLines  150000 non-null  float64
3   age                                   150000 non-null  int64
4   NumberOfTime30-59DaysPastDueNotWorse  150000 non-null  int64
5   DebtRatio                             150000 non-null  float64
6   MonthlyIncome                         120269 non-null  float64
7   NumberOfOpenCreditLinesAndLoans       150000 non-null  int64
8   NumberOfTimes90DaysLate               150000 non-null  int64
9   NumberRealEstateLoansOrLines          150000 non-null  int64
10  NumberOfTime60-89DaysPastDueNotWorse  150000 non-null  int64
11  NumberOfDependents                    146076 non-null  float64
dtypes: float64(4), int64(8)
```

在15万条数据中，月收入属性和家庭成员数目属性存在缺失，缺失数目分别为29731与3924条，其余属性均无缺失值。

在本组实验中，我们统一采用如下的缺失值处理方式：

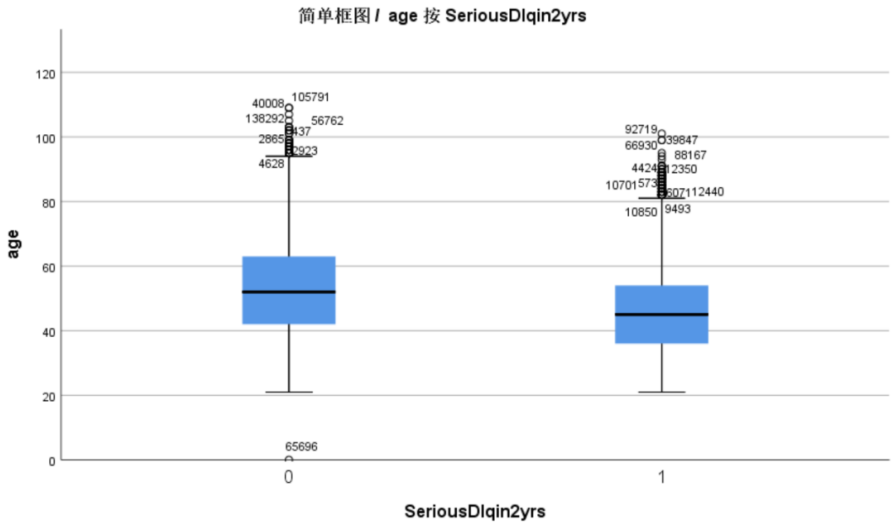
2.1.1 家庭成员数目的处理

由于该属性的缺失量相较于整体而言占比较小，所以我们可以直接采用这150000组数据的众数来填充着3924个缺失数据。经统计，共有86902组数据的家庭成员数目为0，已超过半数。从概率上来讲，把缺失的数据用0来进行填充会有着比较不错的代表性，可以使误差相对。综上，我们采用的方法是用0来填充所有缺失的家庭成员数目值。

2.1.2 月收入的处理

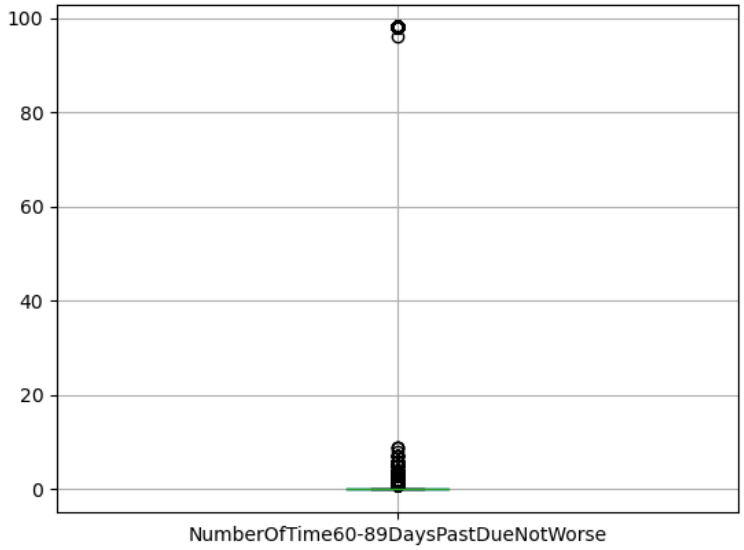
与2.1.1不同，月收入的缺失值达29731个，我们不能用一些简单的样本来替代空缺值，因为即使是一个数据出现的可能性达到了半数以上，在接近30000个缺失数据的基数下，还会有将近15000个数据不甚准确。因此我们最终选用随机森林法来处理数据，即将缺失的数据作为预测值，采用随机森林的计算结果替换缺失值。

2.2 异常值处理



由age箱线图可以看出，年龄存在异常点0。年龄在0——18岁的区间内，对于信用卡的办理问题都是异常的数据。在寻找到对应数据后将其删除。

对于逾期笔数的数据，由下面的箱线图可知存在偏差极大的离群值。经检验，这些数据的三组属性值（逾期30-59天、逾期60-80天，逾期90天）均存在异常值96和98，将异常样本删除。在预测时一旦遇到异常值，我们根据违约概率分别对应到14,4,14。



2.3 对样本不平衡的处理

由于违约与不违约的样本比例相差极大，我们将分别采取过采样与降采样的方法对样本进行处理。具体详见“数据集分析”中的3.1部分。

三、数据集分析

在建立模型之前，需要对数据进行初步的了解，以便建立更好的模型。下面以是否违约及相关性分析为主，进行初步分析。

3.1 对因变量“是否违约”的分析和解读

首先，根据下图我们可以看出，违约人数仅占6.7%。

SeriousDlqin2yrs

		频率	百分比	有效百分比	累积百分比
有效	0	139974	93.3	93.3	93.3
	1	10026	6.7	6.7	100.0
	总计	150000	100.0	100.0	

因变量的数据分布如此不均匀，这就给如何挑选训练的数据集带来了困难。如果仍然保持这个比例来进行训练，以逻辑斯蒂回归模型为例，则会得到如下结果：

分类表[a]

			预测		
			SeriousDlqin2yrs		
	实测		0	1	正确百分比
步骤 1	SeriousDlqin2yrs	0	139165	809	99.4
		1	8749	1277	12.7
	总体百分比				93.6

a. 分界值为 .500

尽管模型的正确率高达93.6%，但是它难以将实际违约的人有效地分辨出来。事实上，

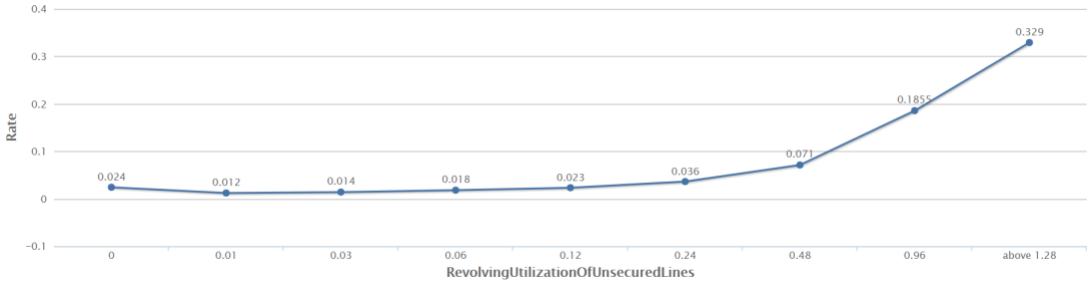
即使是一个无意义的模型：“无论自变量取值是什么，都预测为不违约”，这样也有93.3%的正确率，但它在实际上是毫无作用的。因此，必须对数据集进行调整，消除类别之间的不平衡。由于我国一般将违约风险控制在2%以下，西方国家会控制在4%左右，因此结合实际情况，我们目标是将风险控制在3%以下。以目前的数据来看，在逻辑斯蒂回归模型中，预测不会违约的人中有约6%的人实际上违约了，因此必须增大违约样本的比重，来将违约的人更好的分辨出来。结合将风险控制在3%以下的目标，我们将分别采取过采样和降采样的方法来对数据集加以调整。

3.2 相关性分析

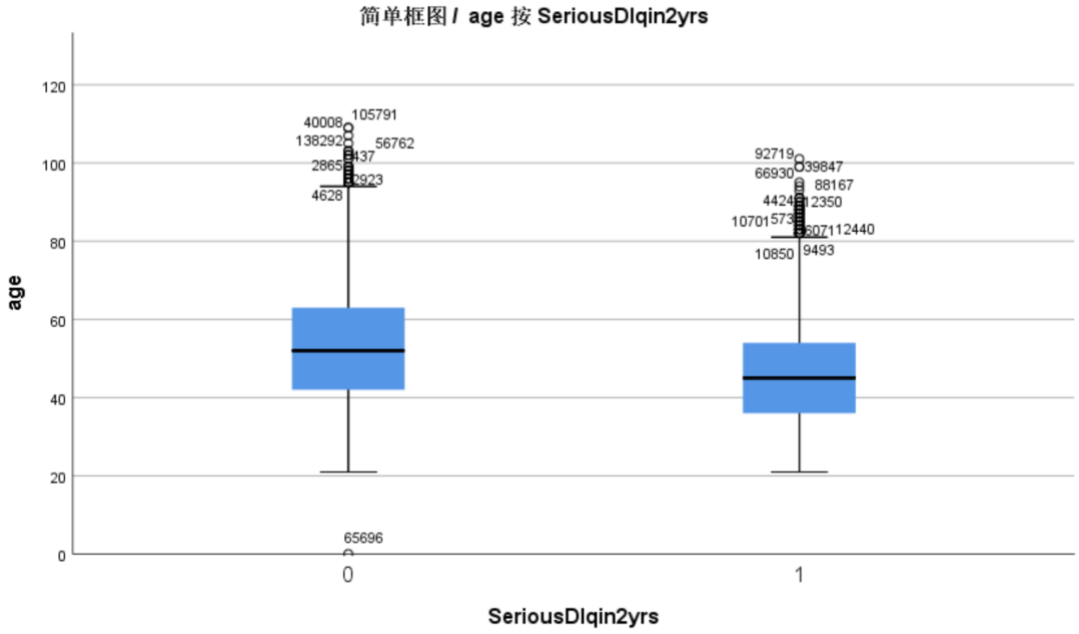
3.2.1 通过数据分布定性分析相关性

在这个部分，我们主要通过观察某个特征关于是否违约的图示来观察该特征与是否违约的关系。关系较为明显的将用箱线图来呈现，箱线图不明显将用违约概率的折线图来进一步展示。

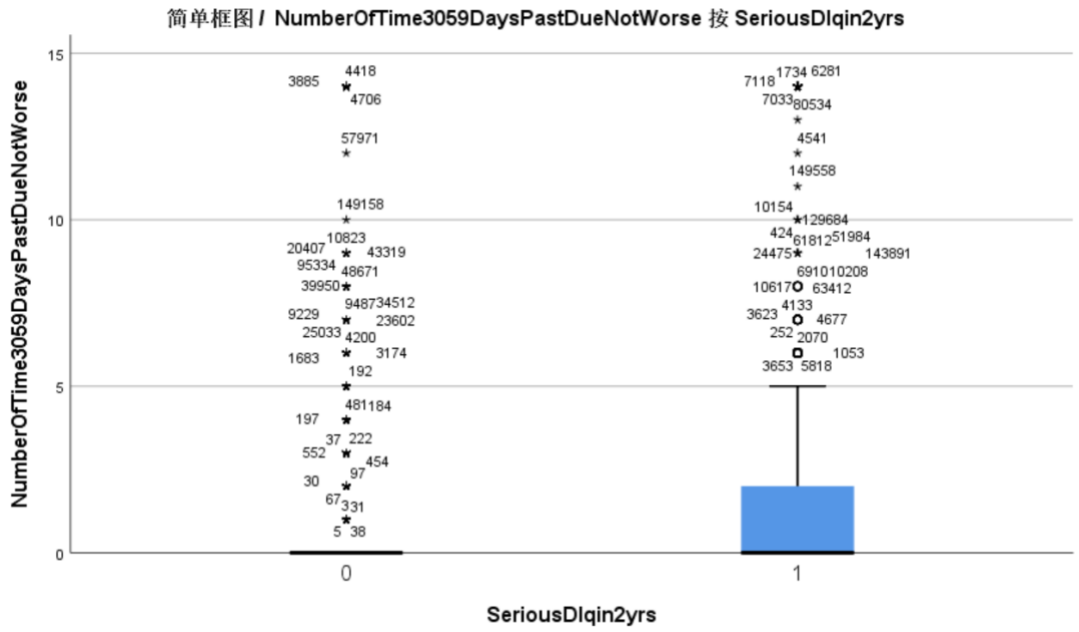
特征RevolvingUtilizationOfUnsecuredLines(下简称Revolving)违约比率的折线图如下，我们可以看出，除了0之外，随着Revolving的增长违约概率也随之增长，说明信用卡总余额和个人信用额度越接近（甚至超出）总信用限制的人越有可能违约。



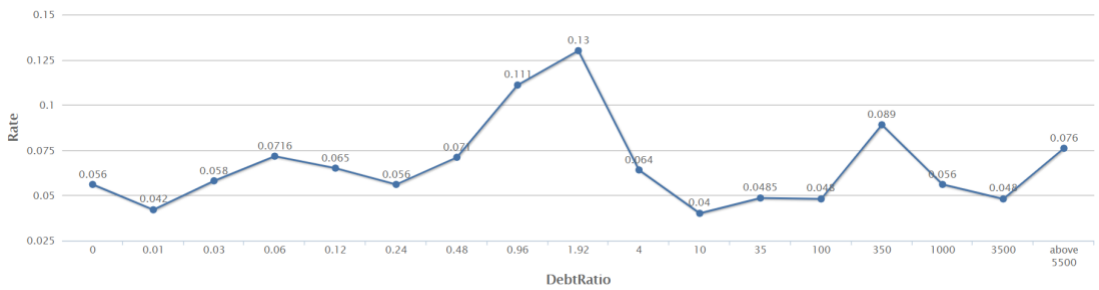
特征age的箱线图如下，可以看出年龄更低的人往往更容易违约，这和现实也基本吻合，年轻人相对更加冲动，收入相对更不稳定，因此违约的可能更高一些。



特征NumberOfTime30-59DaysPastDueNotWorse(下简称30-59Days)的箱线图如下，可以看出几乎全部的未违约者集中在0这个值上，说明一旦出现30-59天欠款逾期的情况，违约概率就会大幅增加。实际情况也是如此，一旦出现逾期情况，那么就很可能违约，这是需要加大关注的一点。

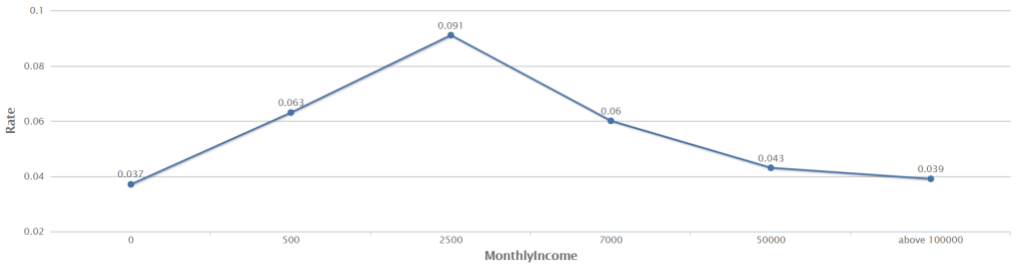


特征DebtRatio违约比率的折线图如下，可以看出该特征和是否违约的关系并不明显。

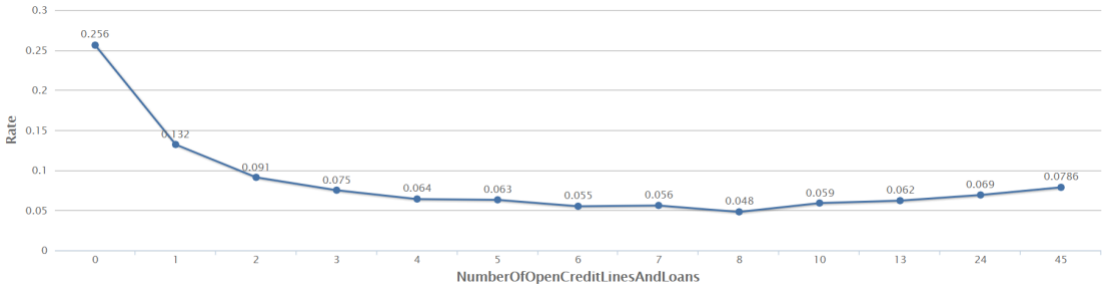


特征月收入违约比率的折线图如下，我们可以看出违约比率随收入的上升先上升后

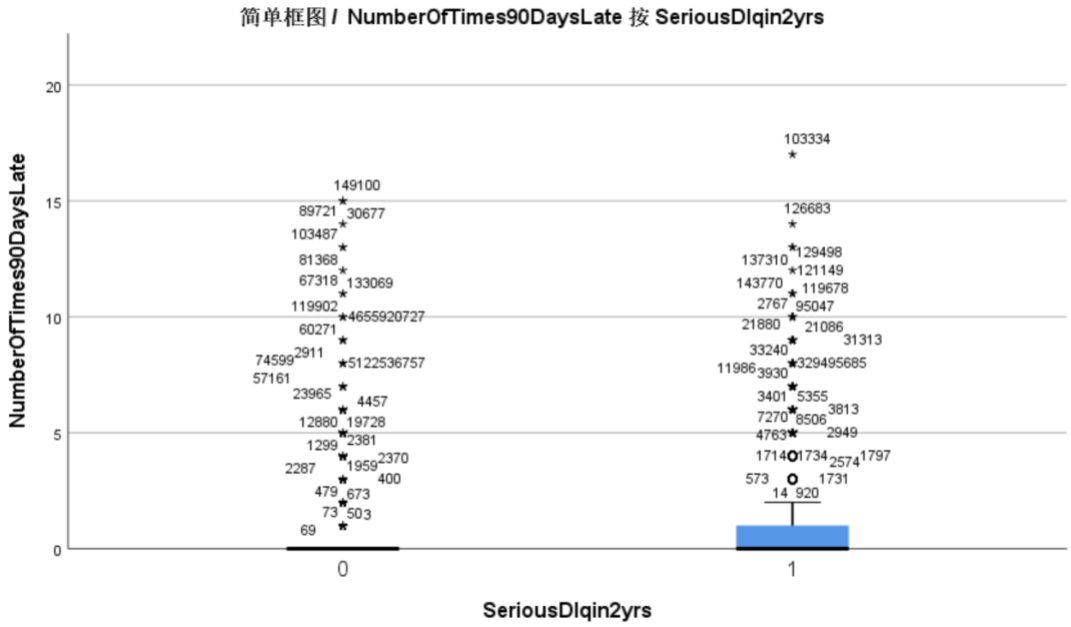
下降。这种情况可能是收入过低甚至没有收入时，对应的人群是老人或青年（事实上在收入不超过500的人中有接近一半的人是老人或青年），他们收入虽然低，但是并不意味着他们贫穷，也不能说明他们更可能违约。在有一定收入之后，违约概率随着收入的增加而降低，这个现象便与常理相符。



特征NumberOfOpenCreditLinesAndLoans(下简称CreditLines)违约比率的折线图如下，我们可以看出在公开贷款和在线信用数量很少时违约比率非常高，之后逐渐下降，当数量过多时有小幅的上升。

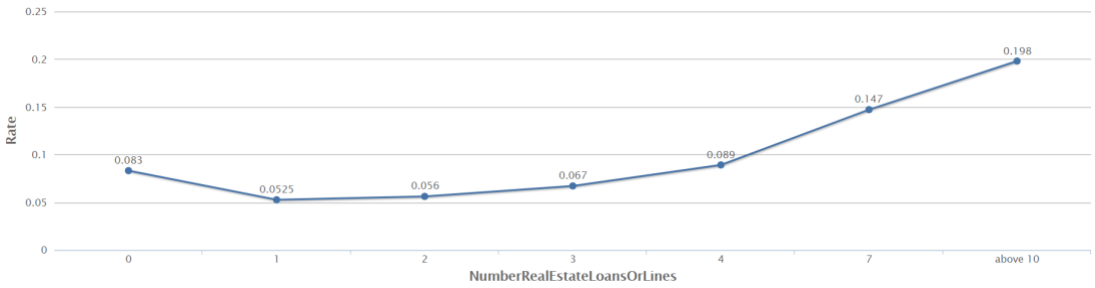


特征NumberOfTimes90DaysLate(下简称90Days)的箱线图如下，我们可以看出几乎全部的未违约者集中在0这个值上，说明90天或以上贷款逾期未还的次数和违约呈比较明显的正相关关系。

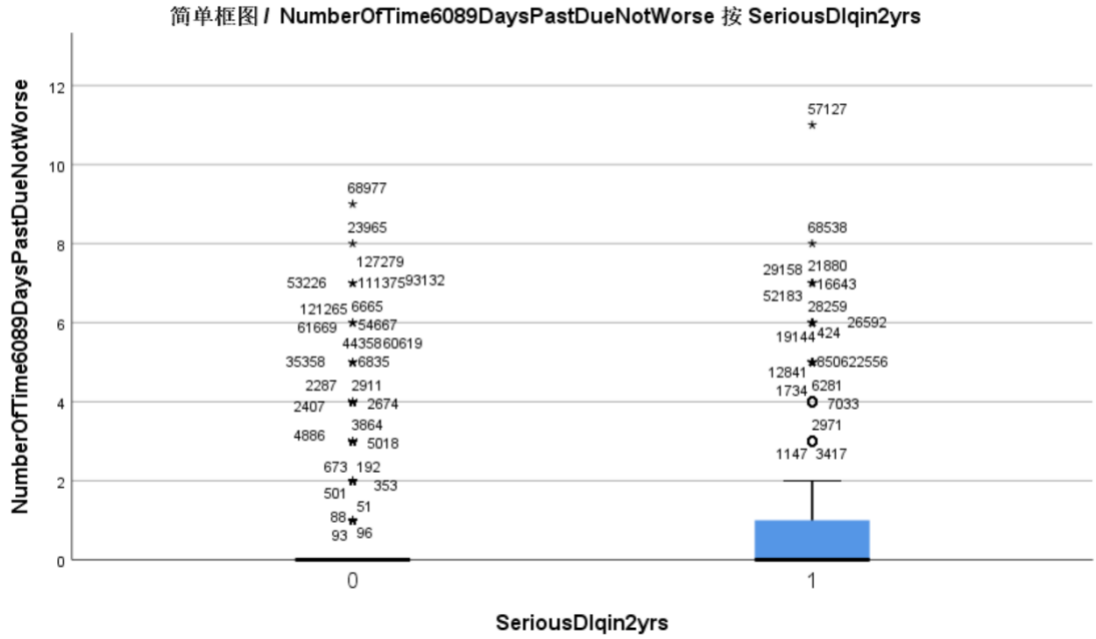


特征NumberRealEstateLoansOrLines(下简称EstateLoans)违约比率的折线图如下，我们可以看出当人们没有抵押或房地产时，就相对更可能违约，而且实际上这样的

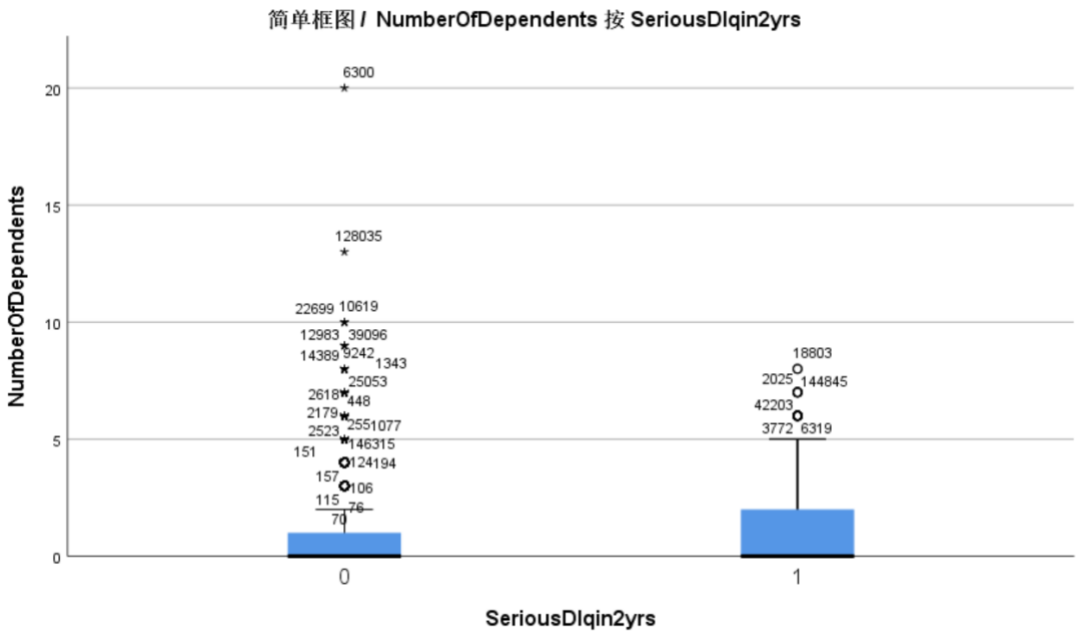
人并不占少数（有超过37%的人）。一旦有了抵押或房地产，那么违约的概率将随抵押和房地产数量增加而增加。



特征NumberOfTime60-89DaysPastDueNotWorse(下简称60-89Days)的箱线图如下，可以看出几乎全部的未违约者集中在0这个值上，和30-59天欠款逾期一样，出现60-89天欠款逾期的情况也意味着违约概率会大幅增加，这也符合常理。



特征NumberOfDependents的箱线图如下，可以看出家庭成员数目越少，违约的概率就越小。这种现象可能是因为家庭成员少，生活压力和负担也就更少，也就有更小的可能违约。



3.2.2定量分析数据相关性

在分析是否违约（SeriousDlqin2yrs）和其它特征的相关性时，采用点二列相关系数；在分析其它特征之间的相关性时，采用皮尔逊相关系数。这是因为除了是否违约是二分变量之外，其余变量均是连续变量。而点二列相关适用于判断连续变量和二分变量之间的相关关系，皮尔逊相关系数是最常用的判断两连续变量之间的相关关系的衡量标准，因此这样选取相关关系的衡量标准是科学合理的。

从下图我们可以看出，是否违约和90Days的相关性最大，与30-59Days和60-89Days也有着较强的正相关关系，另外和年龄也存在着比较明显的负相关关系。这些结论与上面我们定性分析的结果是一致的。其它特征之间的相关性方面，可以看出90Days、30-59Days和60-89Days三者之间的相关性较强。这个现象在实际上也非常符合，30-59天欠款逾期次数、60-89天欠款逾期次数和90天或以上贷款逾期未还的次数很明显有着极强的相关性。另外， CreditLines和EstateLoans也存在着很强的相关性，这个结果也不出意外。

四、模型的建立

4.1 逻辑斯蒂回归模型

4.1.1 模型介绍

逻辑斯蒂回归模型是十分常用的非线性模型，适合处理分类问题。在是否违约这个问题中，可以建立违约概率p和相关因素之间的联系，具体公式可以表示成：

$$ln\frac{p}{1-p} = \beta_0 + \sum_{i=1}^{n}\beta_i x_i.$$

由于最常用的线性回归模型的值域是全体实数，因此并不适合对分类类型的变量建模。而逻辑斯蒂回归模型在线性回归模型的基础上，将值域通过sigmoid函数压缩到

(0,1)之间，便可以解决分类问题了。

4.1.2 参数求解及结果分析

在建立模型之后，就要对影响信用卡是否违约的因素的参数进行求解。首先，我们将每个变量均设为线性关系，即最原始的表达式：$ln\frac{p}{1-p} = \beta_0 + \sum_{i=1}^{n}\beta_i x_i$. 根据极大似然原理和梯度下降的方法，在进行参数求解之后，得到的分类表如下：

分类表[a]

实测			预测 SeriousDlqin2yrs 0	1	正确百分比
步骤 1	SeriousDlqin2yrs	0	103267	36584	73.8
		1	51220	146360	74.1
	总体百分比				74.0

a. 分界值为 .500

我们能够将74.1%的实际违约用户甄别出来，总体的正确率也在74%左右，可见这个模型基本上是有效果的。具体每个变量的系数如下：

方程中的变量

		B	标准误差	瓦尔德	自由度	显著性	Exp(B)	EXP(B) 的 95% 置信区间 下限	上限
步骤 1[a]	RevolvingUtilizationOfUnsecuredLines	.000	.000	2.367	1	.124	1.000	1.000	1.000
	age	-.031	.000	10214.387	1	.000	.969	.969	.970
	NumberOfTime3059DaysPastDueNotWorse	.773	.006	15119.345	1	.000	2.166	2.140	2.193
	DebtRatio	.000	.000	123.165	1	.000	1.000	1.000	1.000
	MonthlyIncome	.000	.000	750.636	1	.000	1.000	1.000	1.000
	NumberOfOpenCreditLinesAndLoans	.014	.001	260.155	1	.000	1.014	1.012	1.016
	NumberOfTimes90DaysLate	1.346	.012	11700.522	1	.000	3.843	3.750	3.938
	NumberRealEstateLoansOrLines	.099	.004	661.977	1	.000	1.104	1.096	1.112
	NumberOfTime6089DaysPastDueNotWorse	1.154	.014	6391.907	1	.000	3.172	3.083	3.263
	NumberOfDependents	.051	.004	193.274	1	.000	1.052	1.045	1.060
	常量	.980	.016	3534.870	1	.000	2.664		

a. 在步骤 1 输入的变量：RevolvingUtilizationOfUnsecuredLines, age, NumberOfTime3059DaysPastDueNotWorse, DebtRatio, MonthlyIncome, NumberOfOpenCreditLinesAndLoans, NumberOfTimes90DaysLate, NumberRealEstateLoansOrLines, NumberOfTime6089DaysPastDueNotWorse, NumberOfDependents。

首先最重要的显著性方面，除了第一个变量均显著，但是第一个变量显著性为0.124，不够显著。结合之前相关性的分析，我们可以知道在该变量取值较小时违约概率较低，取值较大时违约概率会上升，但是Revolving是以指数量级增长的，所以线性拟合的效果会比较差。此外，还有DebtRatio和月收入这两个系数为0。之前已经知道DebtRatio和是否违约相关性不大，因此它的系数是合理的。但是月收入与是否违约呈现先增长后降低的二次曲线的趋势，而且月收入跨度较大，常呈现指数的分级，因此线性拟合月收入与是否违约的关系是否不够准确。

通过上面的分析，我们进行如下处理：将Revolving和月收入进行取对数处理，让得到的值成线性关系。此外，增加月收入对应的二次项这个变量。再次进行参数求解得到如下结果：

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]	V1	.419	.003	20316.879	1	.000	1.521
	V2	-.021	.000	4188.117	1	.000	.979
	V3	.655	.006	10601.976	1	.000	1.924
	V4	.000	.000	117.205	1	.000	1.000
	V5	.372	.012	935.300	1	.000	1.450
	V6	.023	.001	642.719	1	.000	1.023
	V7	1.161	.012	8711.622	1	.000	3.195
	V8	.139	.004	1108.377	1	.000	1.149
	V9	1.050	.015	5157.184	1	.000	2.858
	V10	.041	.004	110.296	1	.000	1.041
	V52	-.038	.001	1441.087	1	.000	.963
	常量	-1.684	.046	1313.333	1	.000	.186

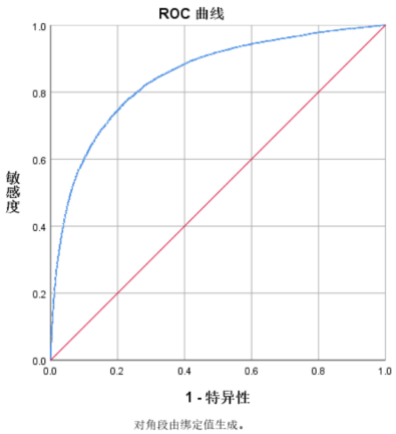
a. 在步骤 1 输入的变量：V1, V2, V3, V4, V5, V6, V7, V8, V9, V10, V52。

这次所有变量的参数都非常显著，这个结果是可信的。下面是对应的分类表和ROC曲线：

分类表[a]

			预测		
			y		
	实测		0	1	正确百分比
步骤 1	y	0	99460	40391	71.1
		1	34600	162980	82.5
	总体百分比				77.8

a. 分界值为 .500



可以看出，这次从违约用户正确识别的比率已经达到了82.5%，相比于之前有了显著的提高。整体的正确率也从74%提高到了77.8%，AUC值达到了0.85。这说明将上面两个变量处理之后还是起到了优化的效果的。

最后，我们得到了是否违约的衡量指标$ln\frac{p}{1-p}=\beta_0+\sum_{i=1}^{11}\beta_i x_i'$，其中$x_i'=x_i(i\neq 1,5,11), x_1'=\begin{cases}\ln(200*x_1^{\frac{1}{4}}+1)(x_1>1)\\ \ln(200*x_1+1)\,(x_1\leq 1)\end{cases}, x_5'=\ln(x_5+1), x_{11}'=(x_5')^2$. 注意到p的分布较为均匀，因此在评分时，我们采取线性的手段将p映射到对应的评分。故分数$S=900-600p$. 经过化简得到最后的评分公式：$S=900-\frac{600t}{1+t}$. 其中$t=e^{\beta_0+\sum_{i=1}^{11}\beta_i x_i'}$, β的具体取值见上表。

4.2 随机森林模型

4.2.1 模型介绍

随机森林模型是一个包含多个决策树的分类器，在构造单个树的过程中，随机选

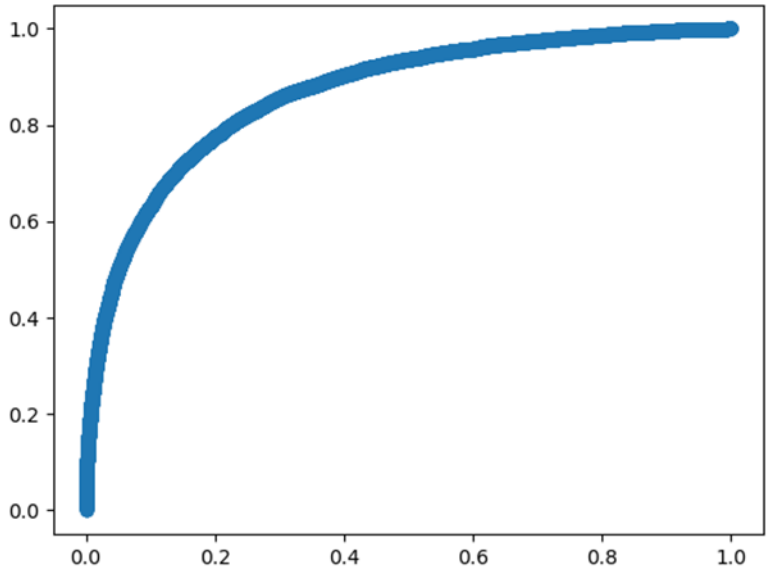
取一些变量或者特征参与树节点的划分，重复多次并保证建立的树之间的独立性。对于每个输入样本，每个决策树对该样本进行判断，得到该样本属于哪一类的结果，并且最终结果由所有决策树中得票最高的类别决定。

随机森林模型具有诸多优点，包括不需要预先做特征选择、并行实现较为容易、通过树的平均降低过拟合的风险和实现简单学习速度较快等。

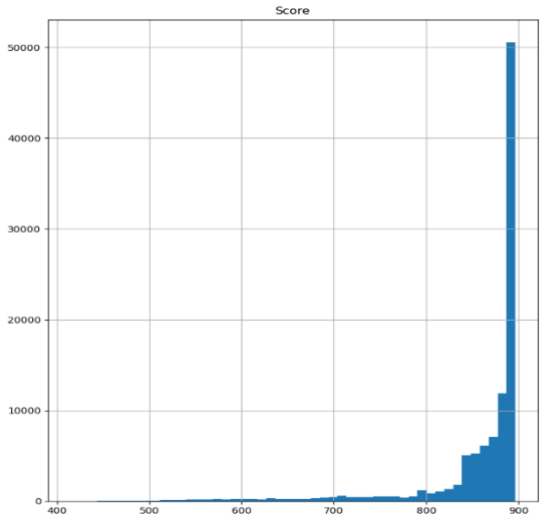
4.2.2 求解与结果分析

使用sklearn中的RandomForestClassifier类。调整参数，使用熵作为决策树的分裂划分标准，规定最大决策树数目。使用cross_val_score进行交叉验证调整参数，并使用两个随机森林分类器的结果进行均值计算。

随机森林分类器的AUC值为0.8676，表明分类器的正确率是较好的。ROC曲线如下图所示。



最终的违约概率需要映射到对应的信用评分。我们采取线性的手段将p映射到对



应的评分。分数$S = 900 - 600p$，得到最终的分数分布如下。

在违约概率的预测中，总体的预测正确率为79.11%，在违约群体和非违约群体中，正确率分别为84.63%和71.12%，满足对于违约群体的重点识别要求。具体的数值如下表。

	0	1	正确百分比
0	99719	40132	0.7112
1	30360	167220	0.8463
总体百分比			0.7788

五、对于过采样和降采样的讨论

前文已经探讨了不同的模型对于实验结果造成的差别。但是，在我们处理问题的过程中，除了构建的模型不同会导致结果的不同外，对于数据的预处理不同也可能会导致最终结果的很大不同，因此我们接下来将要讨论两种不同的预处理对实验结果的影响。

在前面的讨论中，我们已经证明了逻辑斯蒂回归分类法与随机森林分类法在这个问题上具有的不错效果，并且对二者的结果进行了比较。前面的讨论是基于过采样处理的，即把样本扩大化，增加违约样本的个数。然而在许多处理“类不平衡问题”的过程中，人们会采用与之对应的另外一种方法——降采样法。事实表明对于一个150000规模的数据，在一定情况下也会有着相当不错的效果。那么对于这个信用卡建模的问题来讲，是否也会有着这种比较好的效果呢？这就需要我们进行一组对比实验，来研究过采样与降采样对于实验结果的影响。

为了控制变量，我们仍会以逻辑斯蒂回归分类法与随机森林分类法为模型进行实验。与前文提到的模型不同的是，我们会采用降采样法，通过减少不违约样本的数量，使之达到一个与违约样本数量相同的情况，进而我们可以更为准确地训练模型。

我们利用collections包中的Counter函数得到以下结果：

- 原始数据集大小 ： Counter({0: 139974, 1: 10026})
- 降采样后数据集大小 ： Counter({0: 10026, 1: 10026})

可以看到，样本集大小变小，违约样本占总样本数的一半。

本部分我们以ROC曲线与AUC值作为衡量模型好坏的重要指标。

接下来我们从逻辑斯蒂回归模型与随机森林模型两个角度，来分别探究一下降采样数据处理对性能的影响。

5.1逻辑斯蒂回归模型

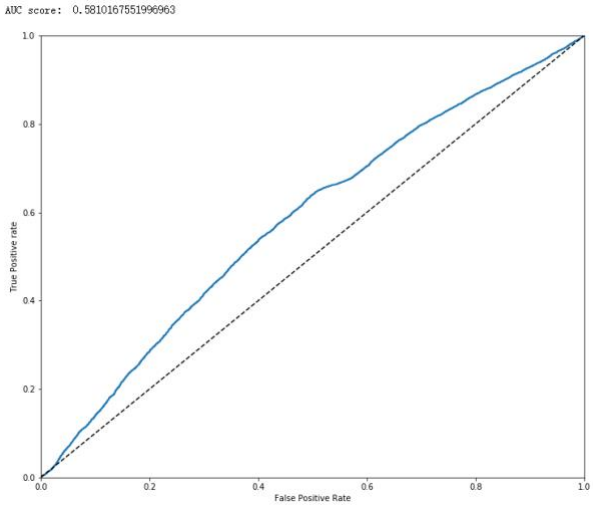
5.1.1模型构建

本部分在模型构建上采取和上述实验4.1.2中一致的模型，且保证了在不对数据

进行任何处理的情况下(以原始的150000条数据进行数据集训练)，所达到的正确率与4.1.2所述模型完全相同，为93.6%，这就保证了两次实验的其余变量均被控制。

5.1.2结果分析

采用降采样法对数据进行处理，得到如下实验结果。



从上述图像可以看到，ROC曲线已经非常接近于主对角线，而AUC值更是只有0.581，远不及同期的过采样模型。实际上，我们一个随机分类器的分类器的AUC值也有0.5.事实证明，对于逻辑斯蒂回归模型，降采样并不是一个明智的选择。

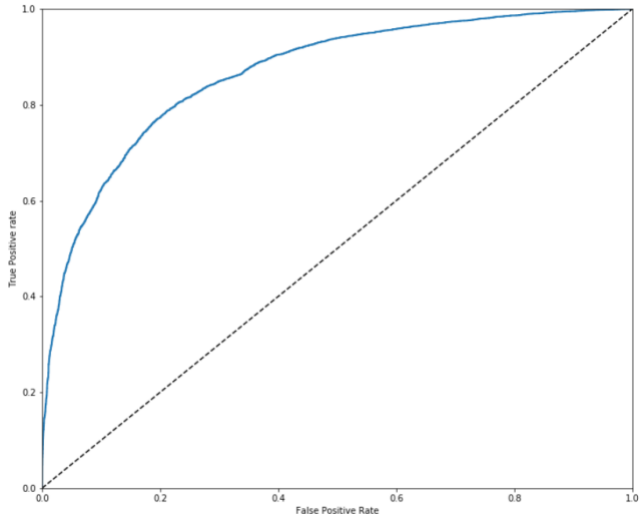
5.2随机森林模型

5.2.1模型构建

本部分在模型构建上采取和上述实验4.2.2中一致的模型，且保证了在不对数据进行任何处理的情况下(以原始的150000条数据进行数据集训练)，所达到的正确率与4.2.2所述模型完全相同。和上面一样，这也保证了两次实验的其余变量均被控制。

5.2.2结果分析

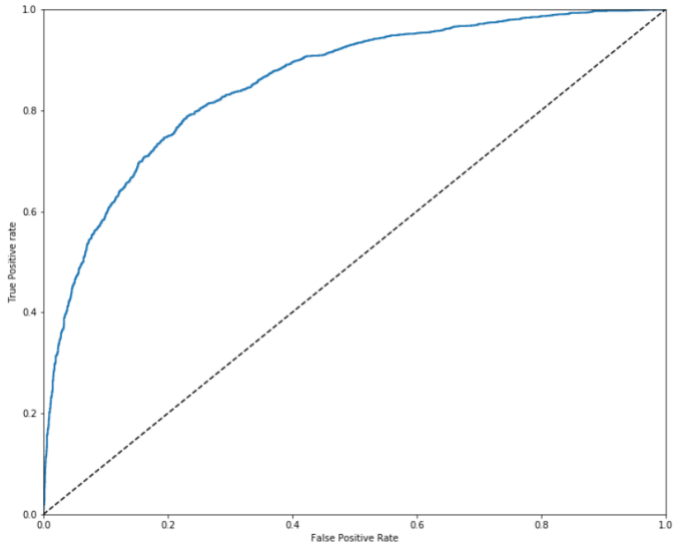
采用降采样法对数据进行处理，得到以下实验结果。其AUC值为0.864.ROC图像如下：



可以看到对于随机森林而言，降采样的处理比逻辑斯蒂回归下的效果要好很多。错

误率明显降低。与4.2.2中的以过采样为数据处理的模型进行比较，可以发现4.2.2中的模型的AUC值为0.8676，二者差别不是很大。这说明对于随机森林分类法，降采样和过采样的结果是近似的，为了验证我们这一结论的正确性，我们进行交叉验证，来确认在随机森林的模型下，降采样的方式没有太多的劣势，或者说与过采样方式结果趋同。

交叉验证所得到的ROC曲线的结果如下：



交叉验证后得到的AUC值为0.853，并没有明显的下降，因此我们得到的结果是：降采样对于随机森林而言，效果与过采样法趋同。

5.3结论

本部分我们以降采样的方式重新构建了两种模型，并基于ROC曲线和AUC值对其分类效果的优劣进行了比较。经比较得出，对于逻辑斯蒂回归模型，降采样有着非常大的缺陷，其结果近似于随机猜测，没有意义；对于随机森林而言，降采样会有着不错的效果，性能上已经十分接近于过采样的处理方式。

目前仅仅做了一个验证，只是想证明对于一个数据不均衡问题，不同的数据处理方式对结果确实会有很大的影响。但其中蕴含的原因，暂时无法得知。希望在进一步的学习过程中，找到这个问题的答案。

六、实验结论

回到我们的实验目的，本次实验我们主要探讨了以下三方面的内容：

- 不同特征的重要程度以及相关性分析
- 不同的模型（逻辑斯蒂回归、随机森林）对于分类结果的影响
- 不同的数据处理方式（降采样、过采样）对于分类结果的影响

6.1相关性分析

6.1.1不同特征的重要程度

不同特征对于违约的重要性可以在与违约的相关性和相应模型求解后得到的系数看出。下面特征后面的括号内为逻辑斯蒂回归模型中的系数。

其中与是否违约呈较为明显的正相关的变量有90Days (1.161)， 60-89Days (1.050)和30-59Days (0.655)。此外，取对数处理之后的特征Revolving (0.419)和EstateLoans (0.139)也发现有着一定的正相关性。与是否违约呈现负相关关系的主要是用户的年龄(-0.021)。此外，随着经过取对数后的月收入的增加，违约频率先上升后下降，对应的二次项和一次项的系数分别是0.372和-0.038。

6.1.2特征间的相关性分析

30-59天欠款逾期次数、60-89天欠款逾期次数和90天或以上贷款逾期未还的次数有着明显的相关性。CreditLines（公开贷款和在线信用数量）和EstateLoans（抵押和房地产数量）也存在着很强的相关性。

6.2不同的模型对于分类结果的影响

经过实验，我们发现，在过采样处理数据的情况下，无论是逻辑斯蒂回归模型还是随机森林模型最终都会得到一个不错的分类结果。其中逻辑斯蒂回归模型违约用户正确识别的比率已经达到了82.5%，总体正确率达到77.8%，违约风险为1.7%（违约风险指在原样本中预测为不违约但实际违约的概率）；随机森林模型在违约群体中正确率为84.63%，总体的预测正确率为79.11%，违约风险为1.5%。这两种模型的预测效果都是比较准确的，而且达到了最初将违约风险控制在3%以下的目标，具有实际意义。如果将二者比较的话，可以发现随机森林模型会略胜一筹，总体正确率与违约样本正确率较逻辑斯蒂回归模型均会高出2个百分点左右。

（注：以下两图分别来源于4.1.2及4.2.2）

分类表[a]

实测		预测 y 0	预测 y 1	正确百分比
步骤 1	y 0	99460	40391	71.1
	1	34600	162980	82.5
总体百分比				77.8

a. 分界值为 .500

	0	1	正确百分比
0	99719	40132	0.7112
1	30360	167220	0.8463
总体百分比			0.7788

6.3过采样与降采样对于分类结果的影响

经过两组控制其他变量的对比实验可以看出，无论是随机森林模型还是逻辑斯蒂回归模型，基于降采样的预处理均不能达到过采样的效果。其中，基于降采样的逻辑斯蒂回归模型正确率更是仅有约50%，大抵相当于一个随即分类器；而基于降采样的随机森林模型效果差强人意，虽比不上过采样的效果，但AUC值与4.2.1中所述模型相差不多，还具有一定的价值。

下表可以概括本次实验模型与采样的核心结论。

	逻辑斯蒂回归模型	随机森林模型
过采样	很好	最好
降采样	无价值	一般