Give Me Some Credit

2020.5.21

CONTENTS

PART ONE

项 目 概 述

一、项目概述

1 问题描述：

- 一种应用统计模型
- 对贷款申请人（信用卡申请人）做风险评估分值
- 规避信用风险
- 申请者评级模型，即对用户进行评估

2 核心问题：

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

PART TWO

数据预处理

0 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)

二、数据预处理

1 缺失值处理

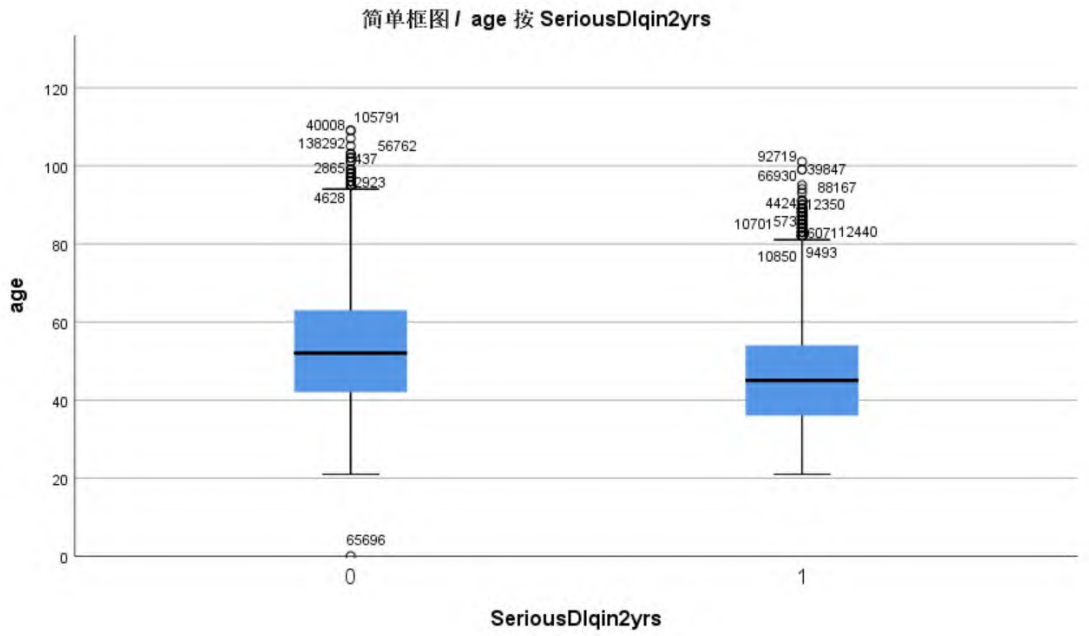
MonthlyIncome属性和NumberOfDependences属性存在缺失
缺失数目分别为29731与3924条，其余属性均无缺失值。

NumberOfDependences的处理：众数0填充
MonthlyIncome的处理：随机森林填充

二、数据预处理

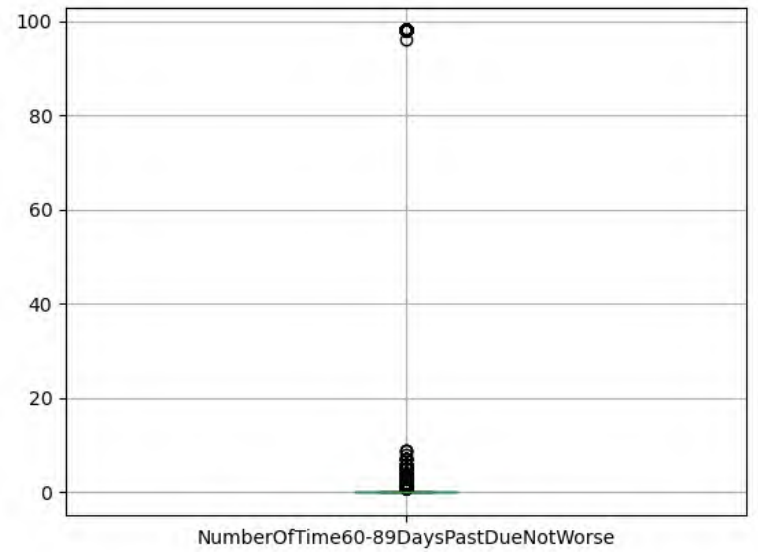
2 异常值处理

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。



2 异常值处理

对于逾期笔数的数据，由箱线图可知
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）
均存在异常情况，将异常行删除。



PART THREE

数 据 集 分 析

三、数据集分析

1 不平衡的数据集

违约人数仅占**6.7%**

目标：将风险控制在**3%**以下

处理方法：过采样、降采样

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	

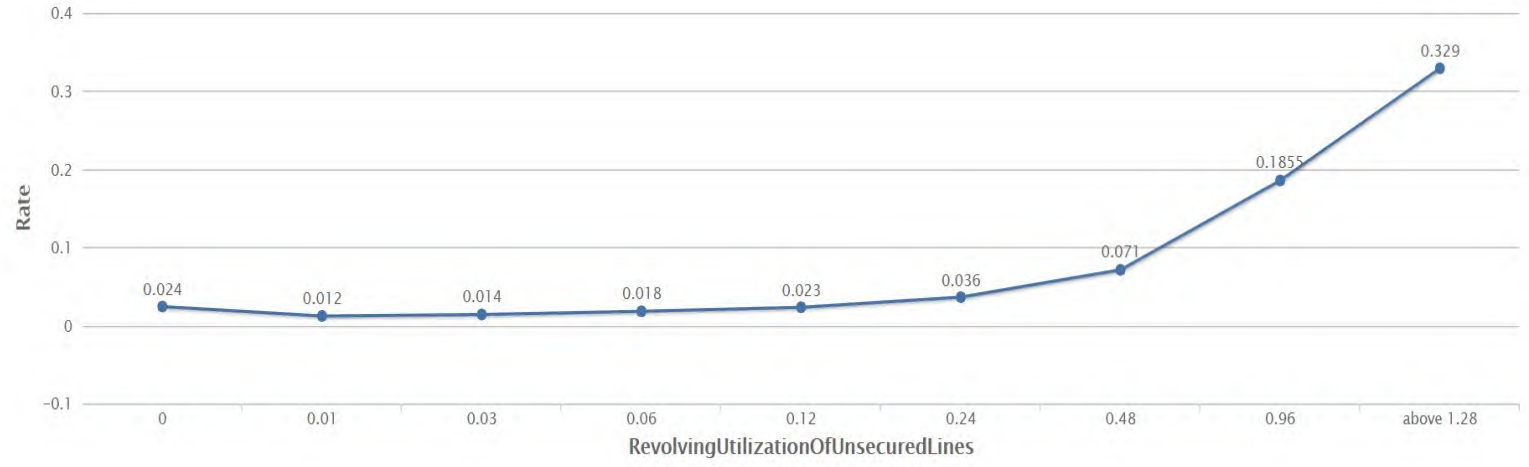
分类表[a]

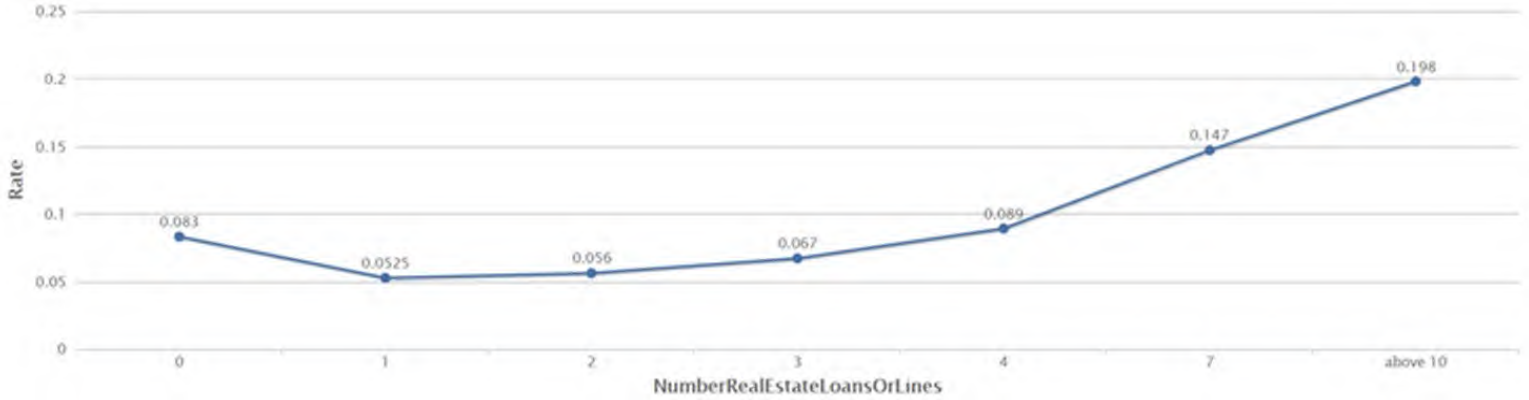
			预测		
			SeriousDlqin2yrs		
	实测		0	1	正确百分比
步骤 1	SeriousDlqin2yrs	0	139165	809	99.4
		1	8749	1277	12.7
	总体百分比				93.6

a. 分界值为 .500

三、数据集分析

2 相关性分析



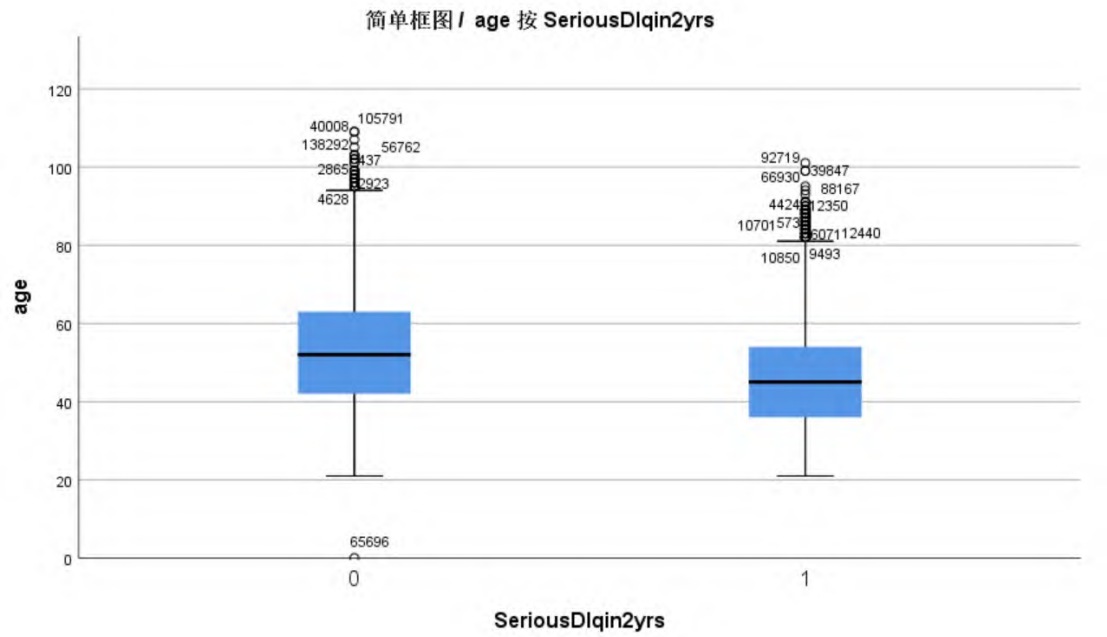


除0外，随着Revolving和抵押数量的增长，违约概率也随之增长

三、数据集分析

2 相关性分析

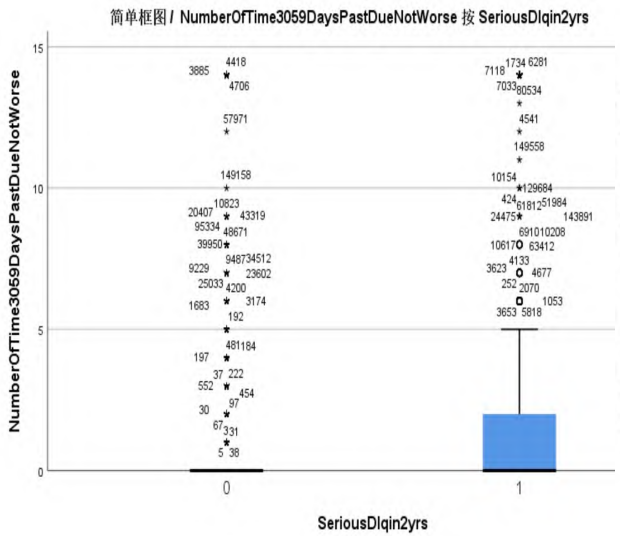
年龄更低的人往往更容易违约

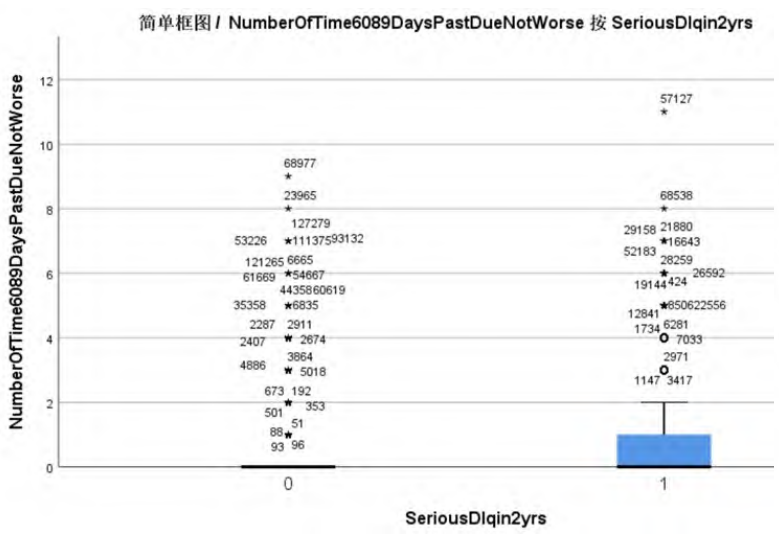


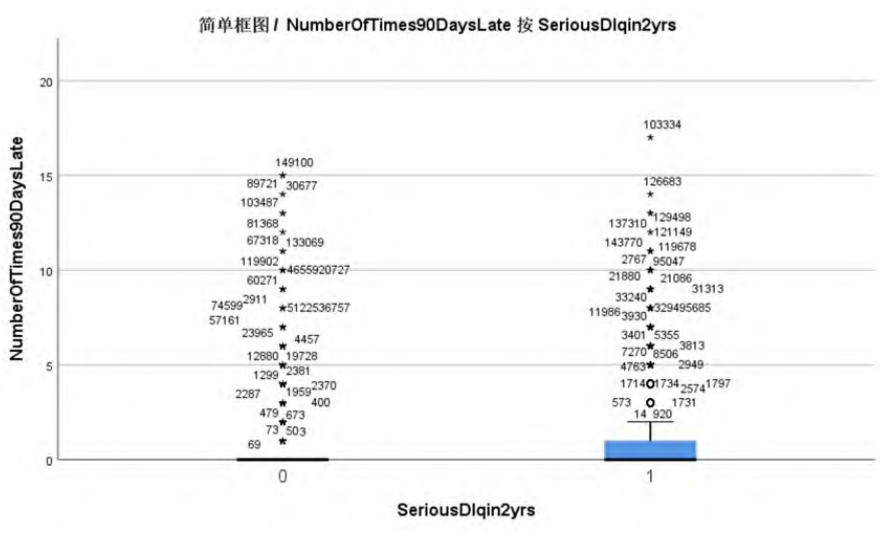
三、数据集分析

2 相关性分析

一旦出现欠款逾期的情况，违约概率就会大幅增加



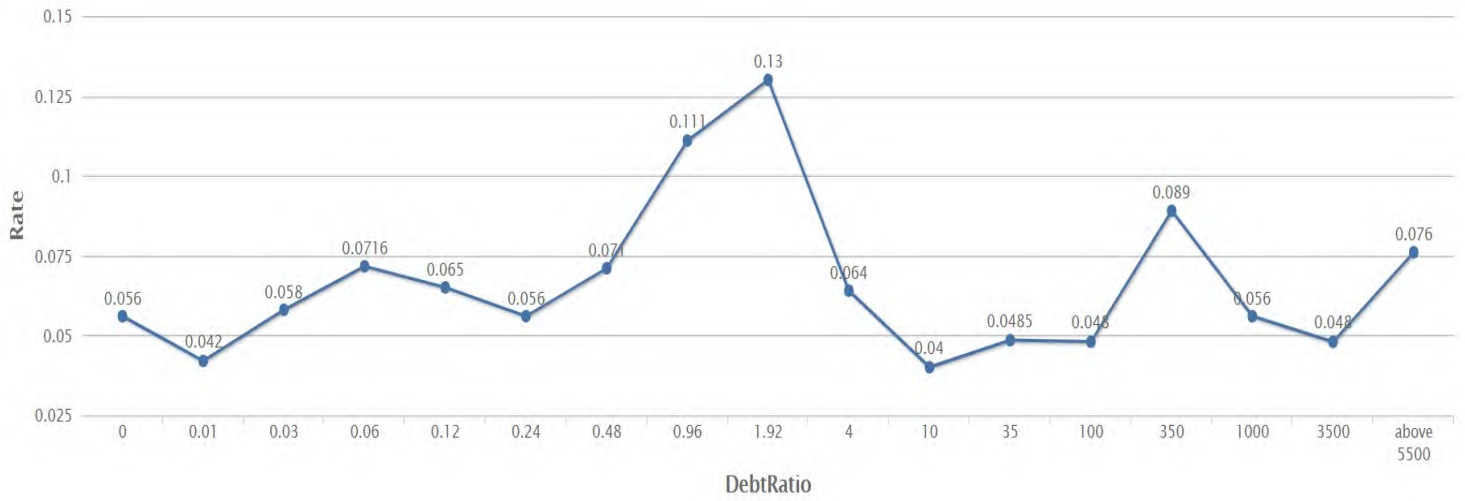




三、数据集分析

2 相关性分析

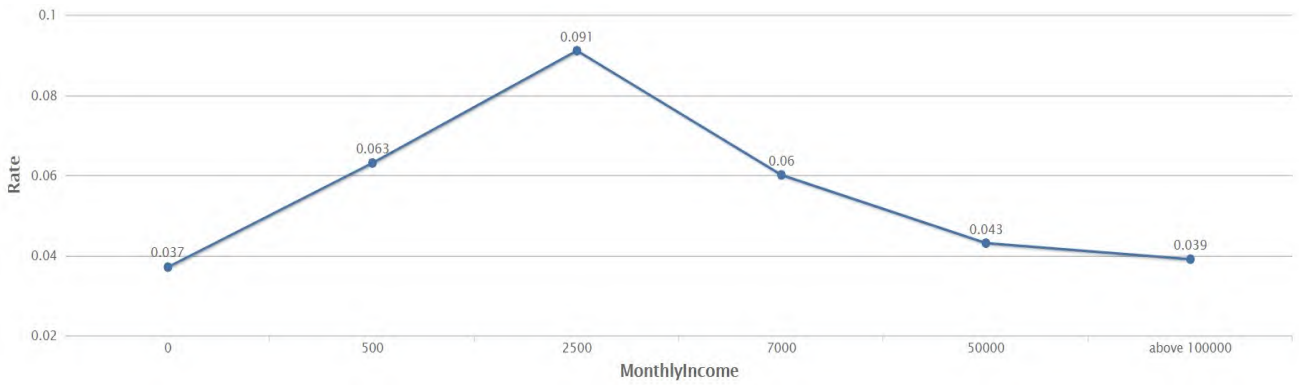
特征DebtRatio违约比率的折线图如下，
可以看出该特征和是否违约的关系并不明显。



三、数据集分析

2 相关性分析

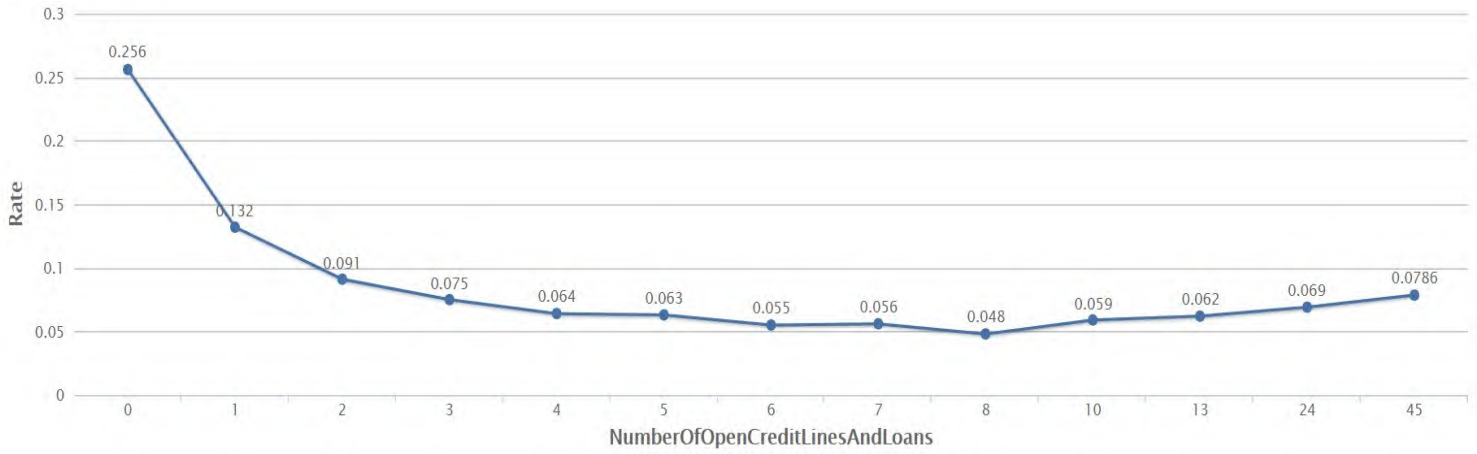
违约比率随收入的增加先上升后下降



三、数据集分析

2 相关性分析

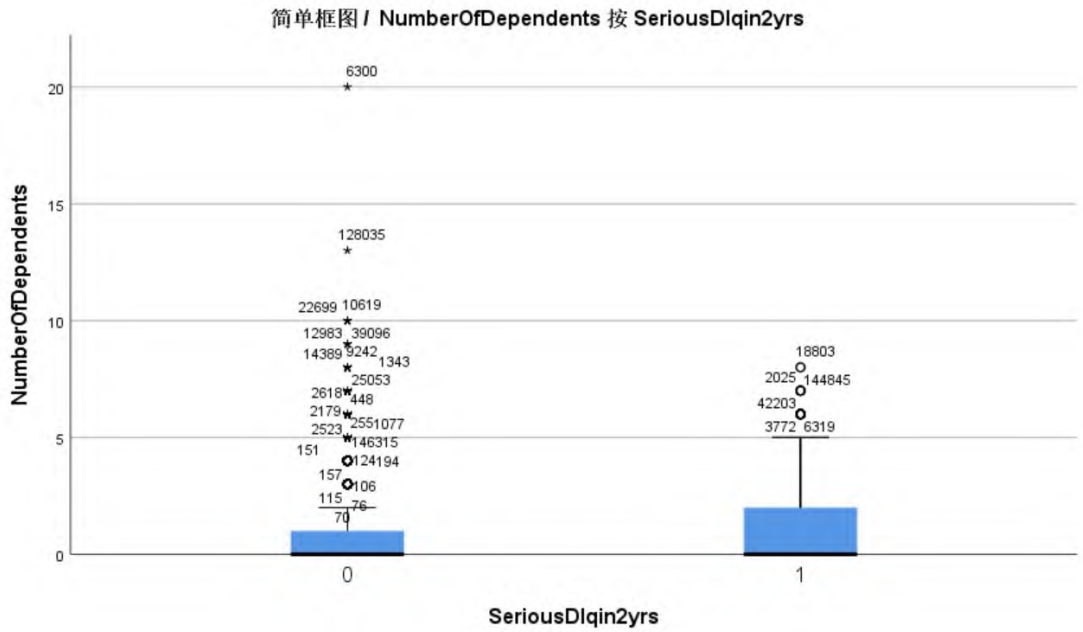
公开贷款和在线信用数量很少时违约比率非常高，之后逐渐下降，当数量过多时有小幅的上升。



三、数据集分析

2 相关性分析

家庭成员数目越少，违约的概率就越小。

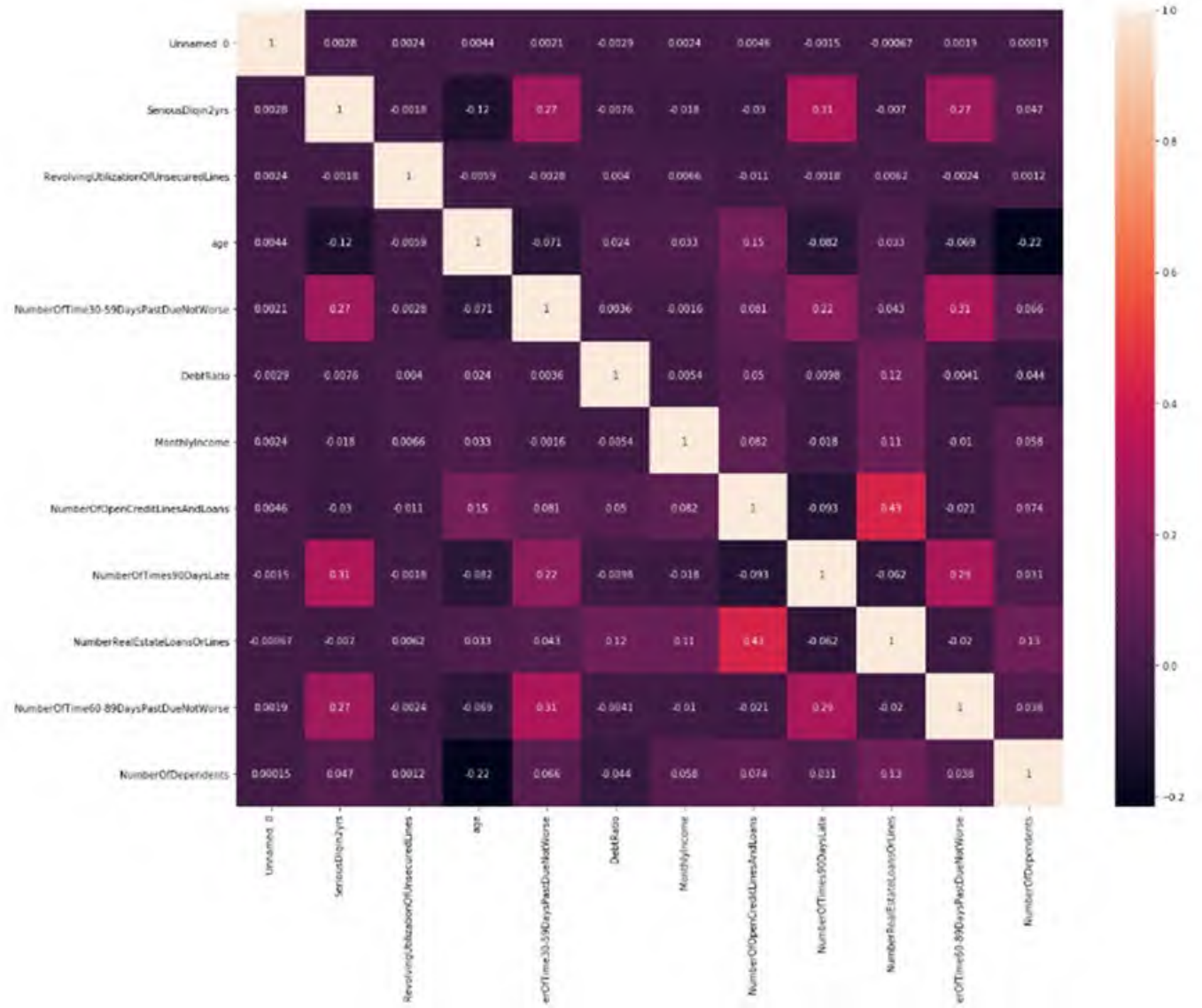


3 定量分析数据相关性

是否违约和其它特征的相关性：点二列相关系数

其它特征之间的相关性：皮尔逊相关系数

3 定量分析数据相关性


Unnamed: 0
SeriousDlqin2yrs
RevolvingUtilizationOfUnsecuredLines
age
NumberOfTime30-59DaysPastDueNotWorse
DebtRatio
MonthlyIncome
NumberOfOpenCreditLinesAndLoans
NumberOfTimes90DaysLate
NumberRealEstateLoansOrLines
NumberOfTime60-89DaysPastDueNotWorse
NumberOfDependents
1.0
0.8
0.6
0.4
0.2
0.0
-0.2

3 定量分析数据相关性

·是否违约和NumberOfTimes90DaysLate的相关性最大

·是否违约与NumberOfTime30-59DaysPastDueNotWorse和NumberOfTime60-89DaysPastDueNotWorse也有着较强的正相关关系

·是否违约和年龄也存在着比较明显的负相关关系。

·NumberOfTimes90DaysLate、NumberOfTime30-59DaysPastDueNotWorse和NumberOfTime60-89DaysPastDueNotWorse三者之间的相关性较强。

·NumberOfOpenCreditLinesAndLoans（公开贷款和在线信用数量）和NumberRealEstateLoansOrLines（抵押和房地产数量）也存在着很强的相关性

PART FOUR

模 型 建 立

四、模型的建立

1 逻辑斯蒂回归模型：

$ln\frac{p}{1-p} = \beta_0 + \sum_{i=1}^{n}\beta_i x_i$，$\beta_i$为待解系数，$x_i$为相关变量的取值.

对影响信用卡是否违约的因素的参数进行求解 准确率：74% 违约样本中的准确率：74.1%

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

方程中的变量

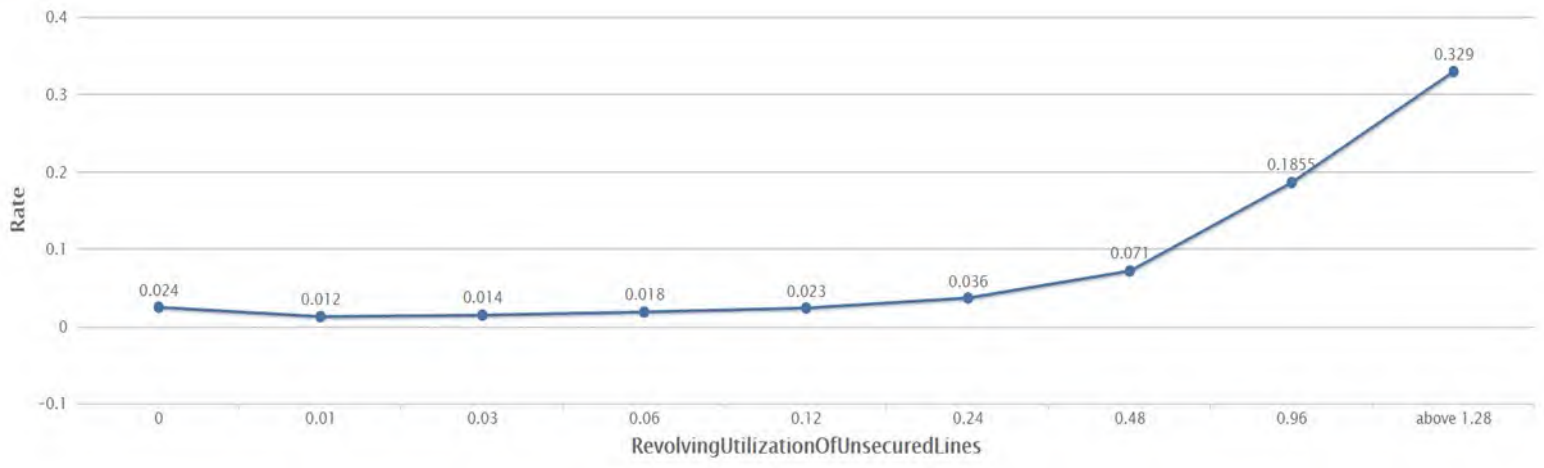
		B	标准误差	瓦尔德	自由度	显著性	Exp(B)	EXP(B) 的 95% 置信区间 下限	上限
步骤 1[a]	RevolvingUtilizationOfUnsecuredLines	.000	.000	2.367	1	.124	1.000	1.000	1.000
	age	-.031	.000	10214.387	1	.000	.969	.969	.970
	NumberOfTime3059DaysPastDueNotWorse	.773	.006	15119.345	1	.000	2.166	2.140	2.193
	DebtRatio	.000	.000	123.165	1	.000	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.

四、模型的建立

1 逻辑斯蒂回归模型

对系数为0的变量进行分析：

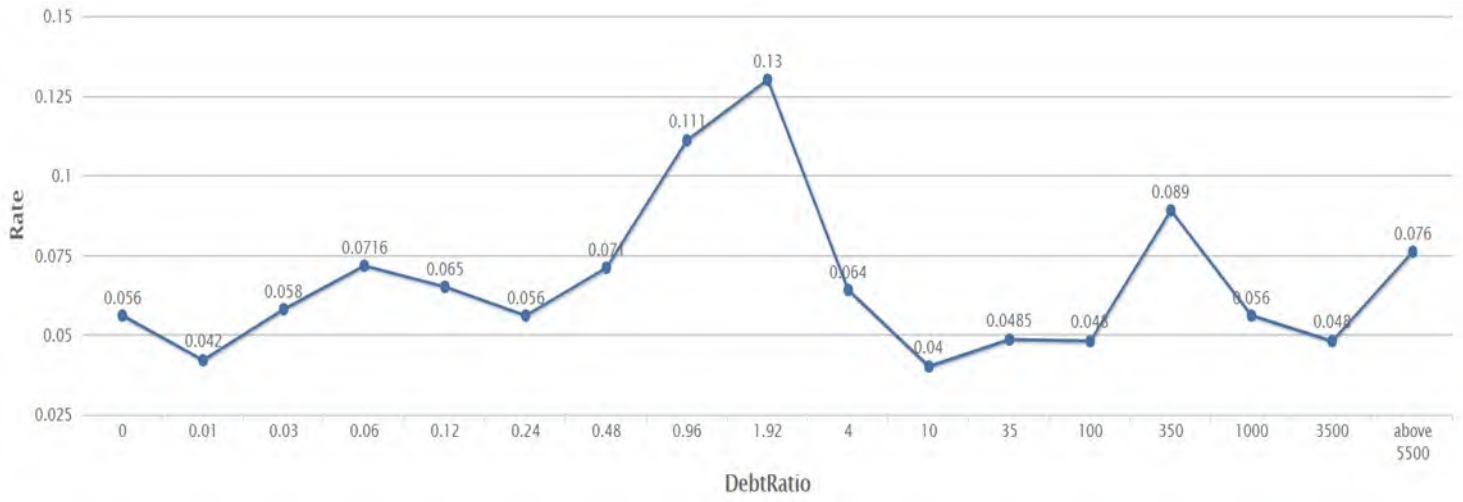


RevolvingUtilizationOfUnsecuredLines：横坐标指数变化，不适合线性求和

四、模型的建立

1 逻辑斯蒂回归模型

对系数为0的变量进行分析：

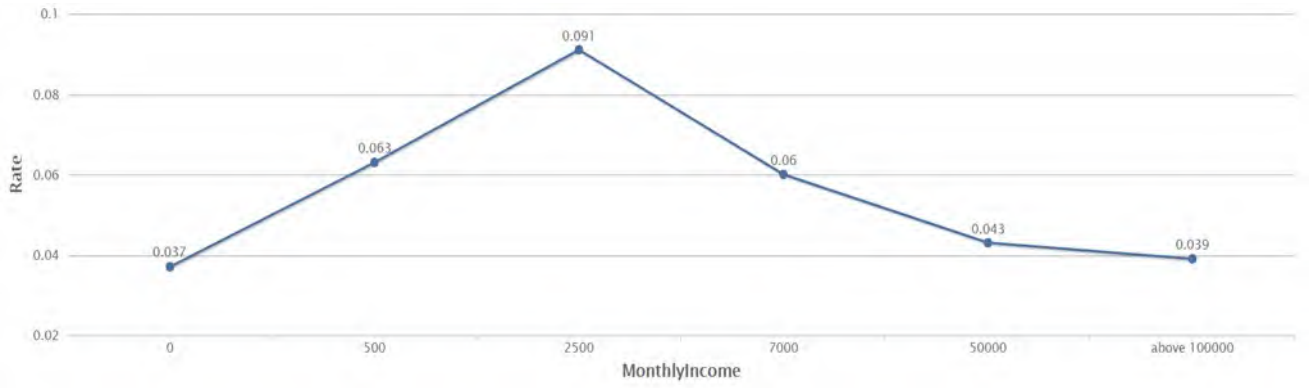


DebtRatio：与是否违约相关性不大

四、模型的建立

1 逻辑斯蒂回归模型

对系数为0的变量进行分析：



MonthlyIncome：先增后降、横坐标指数变化，不适合线性求和

四、模型的建立

1 优化的逻辑斯蒂回归模型：

将RevolvingUtilizationOfUnsecuredLines和MonthlyIncome进行取对数处理，
让自变量取值成线性关系。此外，增加MonthlyIncome对应的二次项这个变量。

准确率：77.8% 违约样本中的准确率：82.5%

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
	y=1	34600	162980	82.5
总体百分比				77.8

a. 分界值为 .500

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。

四、模型的建立

1 逻辑斯蒂回归模型

$$ln\frac{p}{1-p} = \beta_0 + \sum_{i=1}^{11}\beta_i x_i',\ \text{其中} x_i' = x_i(i \neq 1,5,11),$$

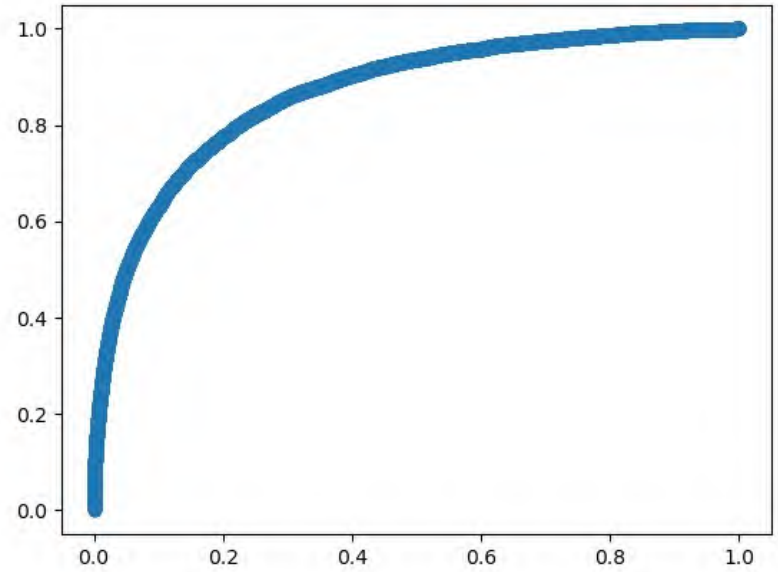
$$x_1' = \begin{cases} \ln(200 * x_1^{\frac{1}{4}} + 1)(x_1 > 1) \\ \ln(200 * x_1 + 1)\ (x_1 \leq 1) \end{cases}, x_5' = \ln(x_5 + 1), x_{11}' = (x_5')^2$$

$$S = 900 - 600p$$

评分公式：$S = 900 - \frac{600t}{1+t}$. 其中$t = e^{\beta_0 + \sum_{i=1}^{11}\beta_i x_i'}$

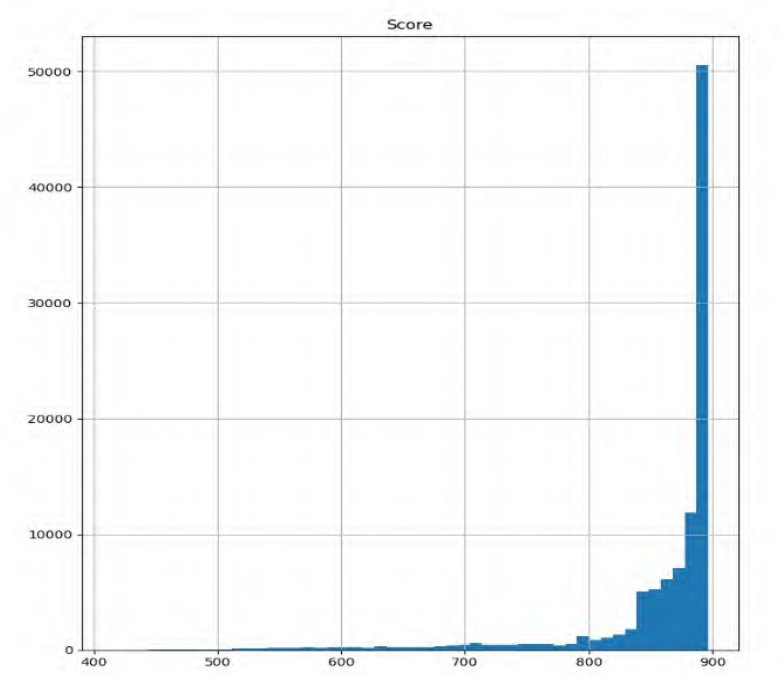
2 随机森林模型：一个包含多个决策树的分类器

随机森林分类器的AUC值为0.8676



2 随机森林模型：一个包含多个决策树的分类器

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右图



2 随机森林模型：一个包含多个决策树的分类器

·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

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

PART FIVE

过采样VS降采样

五、降采样VS过采样

a. 在许多处理“类不平衡问题”的过程中，降采样也是一种不错的选择。

b. 数据集本身有150000的大小，并不是一个小数目，为降采样处理提供了可能。

c. 讨论两种模型下使用降采样都会有什么样的结果

d. 以ROC曲线和AUC值作为重要的衡量指标

五、降采样VS过采样

1 数据的处理与变量的控制

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})

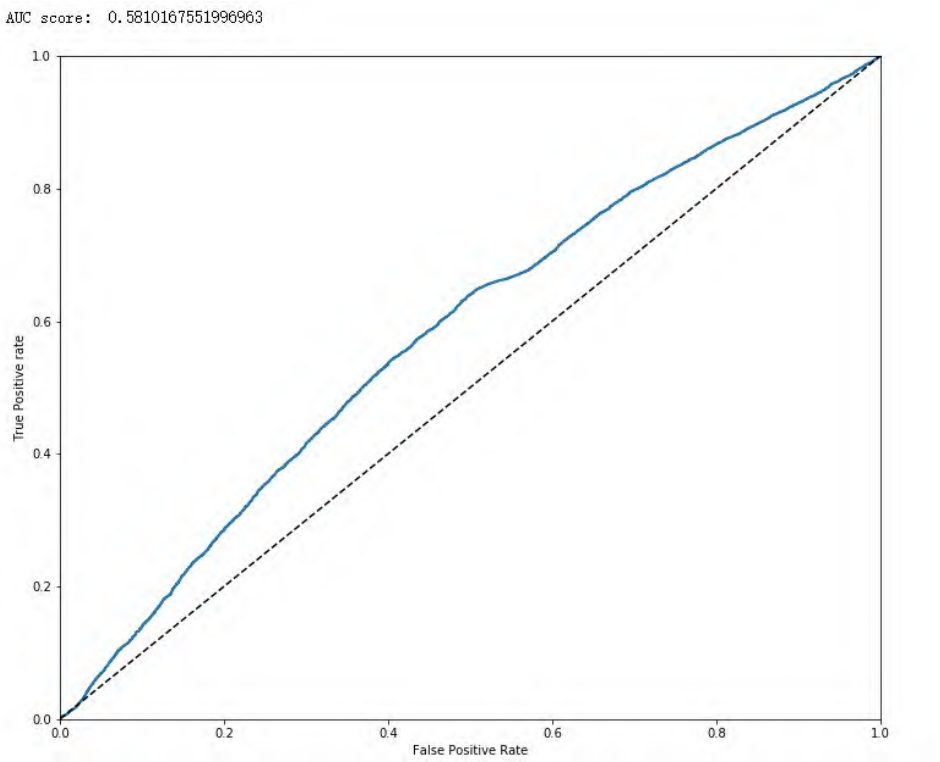
可以看到，样本集大小变小，违约样本占总样本数的一半。

保证在不处理数据的情况下与上述两组实验的正确率相同：控制其他无关变量一定

五、降采样VS过采样

2 逻辑斯蒂回归模型

AUC值为0.581
ROC曲线非常接近于主对角线
与随机分类基本无异！

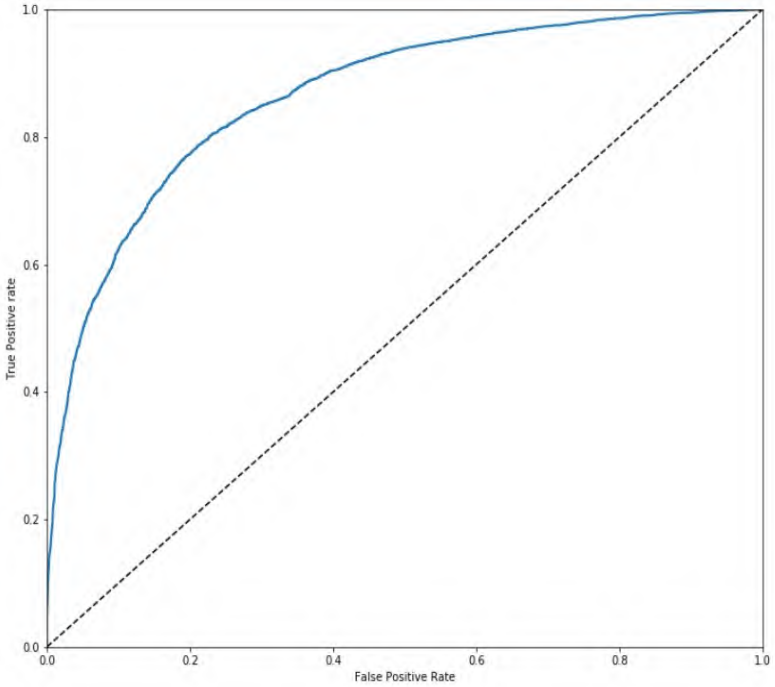


对于逻辑斯蒂回归模型，降采样无效！

五、降采样VS过采样

3 随机森林模型

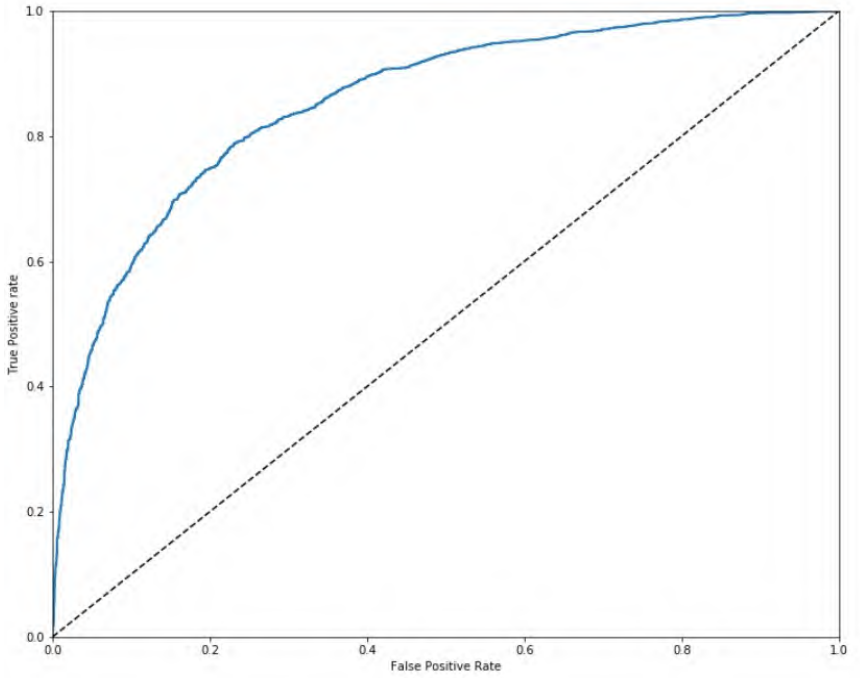
初始AUC值为0.864
已经非常接近于过采样的状态



交叉验证，确定效果是不是真的可以达到和过采样一样

3 随机森林模型

交叉验证得到的AUC值为0.853
并没有很明显的下降



对于随机森林模型，使用降采样的结果很好！

PART SIX

实 验 总 结

六、实验结论

1 相关性分析

·正相关的变量有：

a. 90天逾期次数(1.161)
b. 60-89天逾期次数(1.050)
c. 30-59天逾期次数(0.655)

·负相关的变量有：

年龄

·相关性分析

a. 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
b. 公开贷款和在线信用数量与抵押和房地产数量也存在着很强的相关性

六、实验结论

2 不同的模型及不同的数据处理对于实验结果的影响

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

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

2 不同的模型及不同的数据处理对于实验结果的影响

	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

THANKS !

谢 谢 大 家 ！