信用卡评分建模分析

摘要

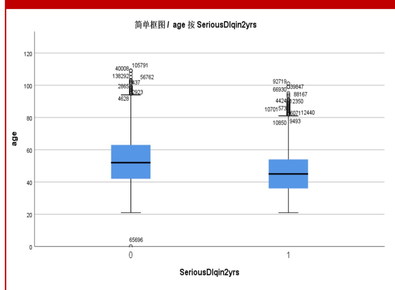
本次实验在对数据进行删除、填充等预处理后，先进行数据的定性相关性分析，并依据不同的相关系数，对数据进行了不同方式的定量分析。之后用过采样的方式调整样本，结合变量优化，使用逻辑斯蒂回归、随机森林等模型进行分类，取得了很好的分类效果。最后对降采样、过采样等法的准确率等效果进行了分析，发现过采样的效果稍优于降采样。

关键词：逻辑斯蒂回归、随机森林、过采样与降采样

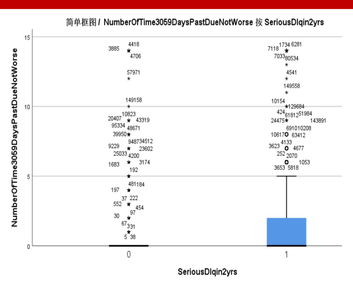
引言

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。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。

数据统计

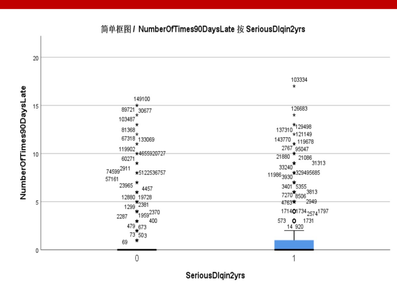


年龄分布

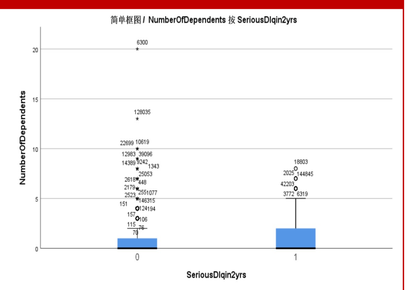


短期逾期次数分布

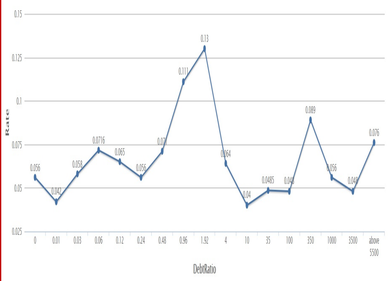
相关系数



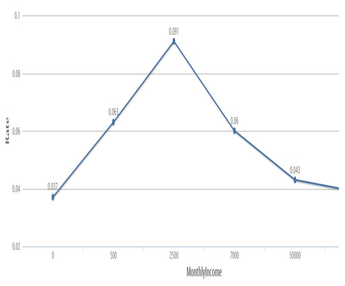
长期逾期次数分布



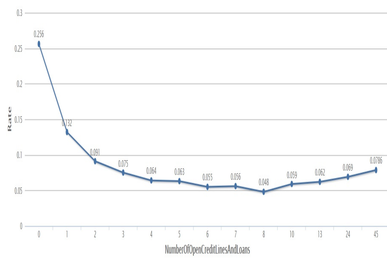
家庭成员数分布



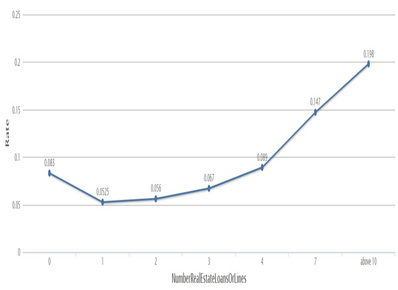
债务比例



月收入



信用额度

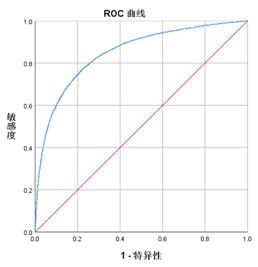


抵押房地产数

模型对比

逻辑斯蒂回归

十分常用的处理分类问题的非线性模型，将最终预测结果表示为$ln\frac{p}{1-p}=\beta_0+\sum_{i=1}^{n}\beta_i x_i$的形式，对某些取值较为特殊的属性进行优化处理，并将概率结果线性映射到最终得分区域。



ROC曲线

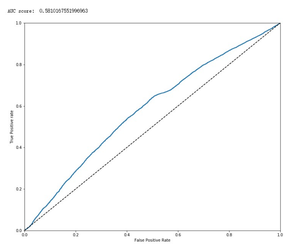
	0	1	百分比
0	99460	40391	71.1
1	34600	162980	82.5
总百分比			77.8

混淆矩阵

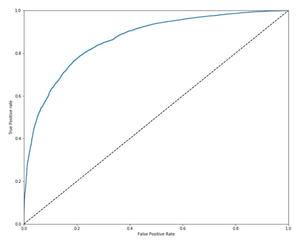
最终的AUC值达到了0.85，表示该分类器具有较好的性能，整体准确率达到了77.8%，其中违约识别率82.5%，满足问题具体要求。

降采样/过采样处理

对于同样的数据，经过降采样法预处理后，得到最终的ROC曲线如下图



逻辑回归

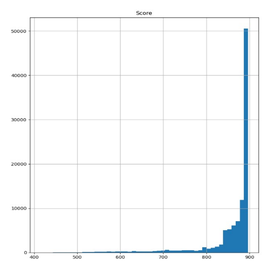


随机森林

由于原预处理使用了过采样方法，由降采样的ROC曲线图与原曲线图对比，逻辑斯蒂回归的性能有着非常明显的下降，甚至接近于随机分类的分类器；而对随机森林来说，0.867的AUC值与使用过采样的值相差不大，仍具有较好的性能。同一处理方式对于不同的模型也有着不同的影响。

随机森林

随机森林模型通过多个特征随机选取的决策树，对其分类结果进行综合，以一定的策略得到综合结果，达到最终的分类目的。随机森林可以一定程度上降低过拟合的风险。



最终分数

	0	1	百分比
0	99719	40132	71.1
1	30360	167220	84.6
总百分比			77.9

混淆矩阵

最终的AUC值达到了0.87，表示该分类器具有较好的性能，整体准确率达到了77.9%，其中违约识别率84.6%，满足问题具体要求。

总结

1、本次实验探讨了不同特征的重要程度以及特征的相关性分析。对是否违约比较重要的几个特征为90天或以上、60-89天、30-59天贷款逾期未还的次数、月收入、信用额度除以总信用限制、年龄、抵押房地产数量。30-59天、60-89天和90天或以上贷款逾期次数有着明显的相关性，公开贷款数量与抵押房地产数量也存在着很强的相关性。

2、本次实验分析了不同的模型对于分类结果的影响。过采样方法的逻辑斯蒂回归与随机森林的准确率均接近80%，并且能够将违约用户有效分辨出来。

3、本次实验对比了不同的数据处理方式对于分类结果的影响，总体上过采样比降采样效果更好一些。

参考文献

[1]王娇. 基于过采样方法的信用卡用户违约预测分析[D].东北师范大学,2019.
[2]Give me some credit | Kaggle notebooks