开元银行卡用户模型分析报告

北京至信普林科技有限公司

二零一八年

目录

[一、 项目背景与目标 1](#_Toc483648686)

[二、 模型变量分析 1](#_Toc483648687)

[（一） 模型变量描述 1](#_Toc483648688)

[（二） 模型变量处理 1](#_Toc483648689)

[三、 建模过程与结果分析 2](#_Toc483648690)

[（一） 算法描述 2](#_Toc483648691)

[（二） 特征重要性分析 3](#_Toc483648692)

[（三） 算法结果评估 4](#_Toc483648693)

[（四） 优化算法与最终结果 5](#_Toc483648694)

[（五） 分数映射 6](#_Toc483648695)

[四、 数据挖掘结果 8](#_Toc483648696)

[（1） 违约情况与总消费金额的关系 11](#_Toc483648705)

[（2） 违约情况与三要素验证的关系 11](#_Toc483648706)

[（3） 违约情况与最近12个月消费金额均值的关系 11](#_Toc483648707)

[（4） 违约情况与月最大消费金额的关系 12](#_Toc483648709)

[（5） 违约情况与在网状态的关系 12](#_Toc483648710)

# 项目背景与目标

# 模型变量分析

1. 模型变量描述

建模数据为：银行卡用户数据。

数据特征：

身份信心验证结果、文化程度、婚姻状况、性别、三要素是否一致、在网时长、在网状态、法院失信被执行人查询、网贷黑名单查询、手机黑名单、不良信息、逾期信息、城市、是否借记卡、首次交易时间、消费总金额、消费总笔数、取现总金额、取现总笔数、近12月没有发生交易的周数占比等。

具体数据描述参见数据质量报告及表*2.2字段描述性统计\_连续型*、2.2*字段描述性统计\_离散型*。

1. 模型变量处理

原始数据字段缺失信息统计如下：



原始数据存在数据缺失严重的问题，根据字段类型，意义采用不同的处理方法，缺失过于严重的字段直接删去，其他根据意义进行填充。

字段选取过程参考的原则是：尽可能赋予属性名和属性值明确的含义；统一多数据源的属性值编码；去除惟一属性；去除重复性；去除可忽略字段；合理选择关联字段。

处理空缺值的方法包括但不限于：忽略该记录；去掉属性；手工填写空缺值；使用默认值；使用属性平均值；使用同类样本平均值；预测最可能的值。

连续值的处理方法：使用归一化处理，消除量纲的影响。

离散型的处理方法：使用条件概率映射和独热编码进行处理。

# 建模过程与结果分析

1. 算法描述

逻辑回归算法：逻辑回归是一种有监督的学习算法，即是有训练集的。在几何层面上，回归的意思，就是我们是否可以找一条线，这条线可以把训练集中的两类数据比较好的分割开来，直线的一侧是类别0中的数据，另一侧是类别1中的数据。这条线就叫做最佳拟合直线，求这条直线的迭代过程就称为回归。然后对于，待分类的一条测试数据，代入拟合直线，看结果在直线的哪一侧，就能得到测试数据的类别，而且可以之间输出变量权重。

随机森林算法：随机森林(Random Forest)由多棵决策树组成，每棵决策树擅长特定人群、重点关注特定变量，然后一起决策，作出最终判断。随机森林在构建每棵决策树时，通过样本和变量两个维度进行随机抽样。随机森林的优点是训练速度快，可以很好地进行并行化，能够处理大规模数据。

XGBoost算法：XGBoost是Extreme Gradient Boosting的缩写，它是Gradient Boosting Machine的一个c++实现，它能够自动利用CPU的多线程进行并行，同时在算法上加以改进提高了精度，在数据挖掘竞赛和工业界都有着广泛应用。Boosting分类器属于集成学习模型，它基本思想是把成百上千个分类准确率较低的树模型组合起来，成为一个准确率很高的模型。这个模型会不断地迭代，每次迭代就生成一颗新的树。对于如何在每一步生成合理的树，有很多的方法，Gradient Boosting Machine是在生成每一棵树的时候采用梯度下降的思想，以之前生成的所有树为基础，向着最小化给定目标函数的方向多走一步。在合理的参数设置下，我们往往要生成一定数量的树才能达到令人满意的准确率。

1. 特征重要性分析

我们利用随机森林输出特征重要度进行分型：

针对银行卡用户数据，共计29个特征。

影响力排名前5的特征依次为：总消费金额、三要素验证、最近12个月消费金额均值、月最大消费金额、在网时长。

1. 算法结果评估

|  |  |  |
| --- | --- | --- |
| **模型** | **AUC** | **KS** |
| **逻辑回归** | 0.63 | 0.18 |
| **随机森林** | 0.66 | 0.25 |
| **XGBoost** | 0.89 | 0.09 |

综合算法表现和人群分布，我们选择随机森林进一步优化。

1. 优化算法与最终结果

通过调参和交叉验证，我们确定模型的最优参数。

最优模型AUC值为0.94，KS为0.74。

我们做出模型在整体样本上的分布图：

由图可知，好人大多分布在50以上，坏人分布大多在50以下，区分明显，建议50为分割线，两部分人群均呈正态分布，符合实际。由于训练样本取自开元通过的人群数据，所以好人分布呈现集中靠右情况，符合实际。

1. 分数映射

我们定义：

* 普通坏账率：违约人数➗总人数
* 加权坏账率：加权违约人数➗加权总人数
* 客户权重：训练模型时被抽中的概率的倒数

根据实际违约情况，我们做出违约分数映射表：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **分数段** | **人数** | **人数区间占比** | **人数累计占比** | **区间坏账率** | **累计坏账率** |
| **95~100** | **3005** | **17.16%** | **17.16%** | **0%** | **0%** |
| **90~95** | **1811** | **10.34%** | **27.51 %** | **0%** | **0%** |
| **85~90** | **1223** | **6.99%** | **34.49%** | **0%** | **0%** |
| **80~85** | **1365** | **7.80%** | **42.29%** | **0.19%** | **0.02%** |
| **75~80** | **3177** | **18.15%** | **60.44%** | **2.08%** | **0.39%** |
| **70~75** | **1500** | **8.57%** | **69.00%** | **1.33%** | **0.51%** |
| **65~70** | **1682** | **9.61%** | **78.61%** | **3.15%** | **0.77%** |
| **60~65** | **1028** | **5.87%** | **84.48%** | **3.11%** | **0.95%** |
| **55~60** | **749** | **4.28%** | **88.76%** | **6.02%** | **1.29%** |
| **50~55** | **787** | **4.50%** | **93.25%** | **12.83%** | **1.86%** |
| **45~50** | **481** | **2.75%** | **96.00%** | **22.04%** | **2.47%** |
| **40~45** | **160** | **0.91%** | **96.92%** | **40.00%** | **2.83%** |
| **35~40** | **138** | **0.79%** | **97.7%** | **60.87%** | **3.31%** |
| **30~33** | **96** | **0.55%** | **98.25%** | **98.96%** | **3.86%** |
| **25~30** | **85** | **0.49%** | **98.74%** | **98.82%** | **4.34%** |
| **20~25** | **74** | **0.42%** | **99.16%** | **100.00%** | **4.76%** |
| **15~20** | **48** | **0.27%** | **99.43%** | **100.00%** | **5.03%** |
| **10~15** | **52** | **0.30%** | **99.73%** | **100.00%** | **5.39%** |
| **5~10** | **29** | **0.17%** | **99.90%** | **100.00%** | **5.55%** |
| **0~5** | **18** | **0.10%** | **100.00%** | **100.00%** | **5.80%** |

根据坏账率映射表，可以结合实际业务设定放贷分数阈值。

# 数据挖掘结果

我们分别作出影响力前五名特征与实际违约情况的分布图进行观察：

（1） 违约情况与总消费金额的关系

（2） 违约情况与三要素验证的关系

（3） 违约情况与最近12个月消费金额均值的关系

（4） 违约情况与月最大消费金额的关系

（5） 违约情况与在网状态的关系