

基于 Logistic 回归的信用评分模型构建与信用规则制定

杨帆 曲阜师范大学

摘要：信用的存在与践行是市场交易的基础，现代征信行业中，以数据挖掘与数据分析为基础的信用评分技术在市场的确定与开拓中起着不可估量的作用。

关键词：Logistic；回归；信用评分；SPSS

中图分类号：F830.479

文献标识码：A

文章编号：1001-828X(2018)018-0011-01

引言

当下互联网金融已蓬勃兴起，呈现出多种多样的业务模式和运行机制。金融机构能够突破时间和地域的约束，在互联网上为有融资需求的客户提供更快捷的金融服务。通过互联网技术，加快业务处理速度，带给用户更好的服务体验。但同时存在着信用风险和用户欺诈等问题，急需通过信用评分模型提高风险控制水平。

一、材料与方法

(一) 建模思路

根据我们所整理的文献资料以及从 FICO 与芝麻信用处所了解的打分标准，总结影响个人履约能力的有以下主要几个要素，包括：

1. 违约历史，包括个人破产，断供房产被拍卖，罚款等情况。
2. 债务负担，即个人杠杆率，包括信用卡信用额度利用率等。
3. 信用种类，客户所用过的信用种类越多，说明其金融知识越丰富，受信方越容易对其信用历史与信用情况做出总结。

根据上述评分标准，综合比较个人信用评分的主要模型，结合所获数据特点，我们首先利用 R 语言原始数据进行清洗，基于处理后的数据确定指标变量，选取 logistic 模型进行建模，再根据 logistic 回归模型的特点，探索一种将履约能力量化的方法，并将此作为最终的信用分数，并据此进行信用评分规则的制定。

(二) 数据处理及指标变量确定

1. 数据处理

由于原始数据相对混乱且数量巨大，同时存在大量缺失值，我们首先利用 R 语言对原始数据进行清洗，并将所有表根据整理后的变量合为一张表，确认每位用户的相关的指标情况。

2. 指标变量确定

经过整理指标变量，总结了个人特征变量两个，包括婚姻状况 MARRY 以及教育程度 EDU；经济特征变量七个，包括贷款金额 LA、未结清贷款金额 ULA、未销户贷记卡合同金额 DCA、未销户准贷记卡合同金额 PDCA、贷款逾期笔数 ONLA、贷记卡逾期账户数 ONDC 以及准贷记卡逾期账户数 ONPDC。并对指标变量进行量化处理。

虚拟变量对应的属性为“真”时选择“1”，为“假”时选择“0”。以“教育程度”为例，当样本数据中“教育程度”指标为初等教育时，虚拟变量“教育 0”取 1，其他虚拟变量取 0；“教育程度”指标为中等教育时，虚拟变量“教育 1”取 1，其他虚拟变量取 0；“教育程度”指标为高等教育时，虚拟变量“教育 0”，“教育 1”都取 0。

我们以用户“183728”为例（变量对应的属性为“真”时选择“1”，为“假”时选择“0”），该用户已婚，接受过中等教育，有 30 万~50 万的贷款金额，目前还有不到 5 万的贷款金额未结清，有不到 5 万的未销户贷记卡合同金额，同时还有 1 到 3 笔贷款逾期笔数。

二、结果与分析

(一) 信用评分模型的构建

1. 模型选择

回归分析中，因变量 Y 有两种情形，即定量变量和定性变量，而 logistic 回归模型是在于研究因变量非连续变量情况下的分析模型。logistic 回归模型作为经典的信用评分模型，将实数域上的输出映射到 [0,1] 上，且具有较强的可解释性。

2. Logistic 回归模型的构建与求解

logistic 回归的结果可以看出，教育程度 EDU、贷款金额 LA、未销户贷记卡合同金额 DCA、贷记卡逾期账户数 ONDC 和准贷记卡逾期账户数 ONPDC 这五个变量的 p 值均小于 0.05，可以认为教育程度、贷款金额、未销户贷记卡合同金额、贷记卡逾期账户数和准贷记卡逾期账户数与因变量之间的线性关系在统计上是显著的，保留五个变量在模型中。

最终采用前进法，将显著的指标变量未销户贷记卡合同金额 DCA、贷款金额 LA、贷记卡逾期账户数 ONDC、教育程度 EDU、准贷记卡逾期账户数 ONPDC 纳入回归方程。

(二) 信用评分规则的制定

1. 指标变量得分计算

根据评分模型所求得测试集的信用分数最小值为 243，而最大值为 321，将客户的信用得分情况与其原有资料相对比，信用得分在 300 分以上的客户，基本为履约客户，很少出现履约或逾期不还现象；而得分在 280 分以下的银行客户，其在银行信用记录上多次出现准贷记卡逾期账户或贷记卡逾期账户，属于所界定的违约客户。

2. 信用评分规则的制定

根据评分公式计算客户信用得分，得分区间为 [240,328]。随着所得分数的上升，客户的信用随之提升，违约概率相应减小，反之，信用分数越低，客户违约概率越大。结合训练集数据信息，信用得分在 300 分以上的客户，信用评估较好，基本为履约客户；而得分在 280 分以下的银行客户，其违约概率较大，此时银行应谨慎考虑是否继续放贷。

三、讨论

(一) 结论

为了有效控制和防范信贷风险，银行必须充分利用已经掌握的数据，对借款人利用信用评分模型进行准确的风险评估。信用评分模型，选择了五个虚拟变量，制定评分规则，计算信用得分，银行可以根据得分高低，评估借款人的信用风险，进而选择是否放贷，从而有效降低风险，实现效益最大化。但是，信用评分模型也存在缺陷。

(二) 建议

为使得银行降低个人信贷风险，保持经济健康稳定发展，一方面，银行应当建立全面而精准的风险管理体系，不断完善客户信用档案，提高信用评估模型的准确性、稳定性以及解释性；另一方面，在使用信用评分模型时，要关注其时效性，及时更新模型，在具体使用过程中，可以凭借审批经验，对信用等级的分数划分进行及时调整，谨慎评估后做出是否放贷的决定，如此才能在最大程度上保证授信方与受信方的利益，并且开拓更加广泛的市场。

参考文献：

- [1] 廖国民, 涂稳华, 宁静. 基于 Logistic 模型的个人消费信贷风险评估 [J]. 广东外语外贸大学学报, 2013(5):27-33.
- [2] 张国政, 陈维煌, 刘呈辉. 基于 Logistic 模型的商业银行个人消费信贷风险评估研究 [J]. 金融理论与实践, 2015(3):53-57.
- [3] 王磊, 范超, 解明明. 数据挖掘模型在小企业信用评分领域的应用 [J]. 统计研究, 2014,31(10):89-98.
- [4] 石庆焱, 靳云汇. 个人信用评分的主要模型与方法综述 [J]. 统计研究, 2003(8):36-39.
- [5] 姜明辉, 徐佩, 任潇. 个人信用评分模型的发展及优化算法分析 [J]. 哈尔滨工业大学, 2015,47(5):40-45.