**项目计划书**

**一、项目背景**

巴西银行(Banco do Brasil)，是巴西最大的银行，也是拉丁美洲最大的银行。它成立于1808年，总行设在巴西利亚，是巴西国有银行。巴西银行在巴西境内有近4万家网点，业务涉及银行业的所有领域。服务的行业主要包括：农业，飞机，通讯，[计算机](http://baike.baidu.com/view/3314.htm)，基础建设，数据处理，工业医药，石油等。巴西银行同时是一家国际化银行，其由于准确的市场把握能力使得其业务范围遍及世界40多个国家，拥有庞大的海外分支机构网，在所有的世界金融中心都有巴西银行的身影。

信用卡业务是巴西银行主要的零售业务之一，也是其个人业务的主要利润来源之一。但自2008年国际金融危机之后，信用卡业务的违约率大幅上升，从而对信用卡业务的利润带来严重威胁。2014年1月份，信用卡风险控制部门（以下简称“部门”）决定开发一张信用卡的申请评分卡，从而更好的控制信用卡业务的风险。评分卡建成之后将在该行的信用卡审批系统NBSM中嵌入实施。

**二、数据**

银行相关部门决定选取2011年10月份到12月份这3个月申请信用卡的用户用来建模。在这三个月中共有150123人申请巴西银行的信用卡，其中2345人中途放弃了申请，57283人被拒绝，剩下90495人获得批准。

用来建模的数据主要分为两部分，一部分是客户申请是提供的一些基本信息（性别、年龄、收入等等），数据见modeling\_demo.sas7bdat。另一部分是巴西银行从巴西信用局获取的客户在申请时的信用记录（贷款和信用卡总数、债务比、过去两年的逾期记录等等），数据见modeling\_buro.sas7bdat。

其中，id, age, dependent, num\_line, num\_loan, del\_30, del\_60, del\_90这八个变量为integer（整型），gender变量为categorial（分类变量），debtratio变量和utilization变量为percentage（百分比类型），income变量是real（双精度浮点型），bad变量是Y/N类型。

在做数据处理时，首先把modeling\_buro数据集里id缺失的观测删掉；然后将modeling\_demo和modeling\_buro 以id为键进行合并；对income变量的缺失值做均值插补；最后把del\_30变量为96或98的数据删掉，此时del\_60和del\_90里等于96或98的异常也被删掉了。

观察窗口和表现窗口均为24个月，“坏”客户定义在在表现窗口中出现过90天及以上的逾期。有394个客户在表现窗口内丢卡或者死亡，因此剩下90101个客户具有表现信息。

2012年1月份的数据被用来作为验证样本，具体数据见validation.sas7bdat。2014年1月份的数据被用来做PSI检验，具体数据见psi.sas7bdat。

Waterfall如图1所示，Vintage Curve如图2所示，Roll Over Rate如表1所示

观测点上所有信用卡申请者

被拒绝的客户

被批准的客户

撤销申请的客户

通过交叉验证获得的客户

在表现窗口内死亡或者卡丢失的客户

其他客户

客户号遗失的客户

图1

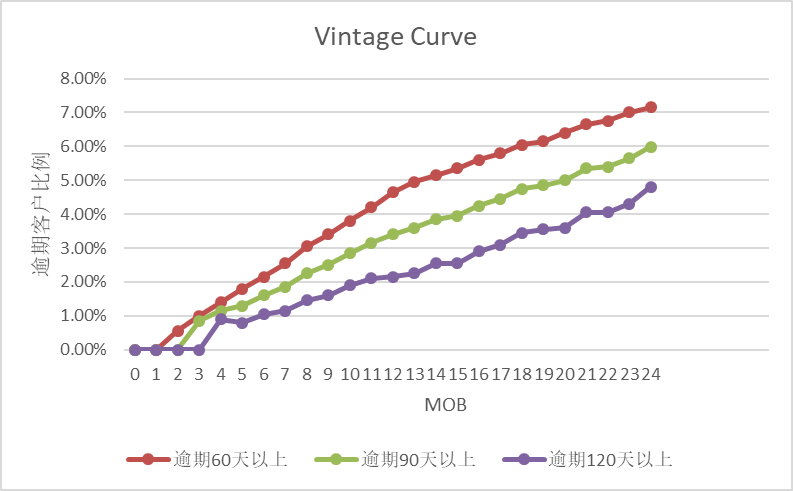


图2

表1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 2013年10月 | 2013年11月 | 2013年12月 | 2014年1月 |
| X 到 30+ | 13.1% | 15.1% | 14.2% | 13.9% |
| 30+ 到 60+ | 32.1% | 30.1% | 35.2% | 33.3% |
| 60+ 到 90+ | 54.1% | 50.1% | 45.2% | 53.3% |
| 90+ 到 120+ | 82.4% | 80.4% | 85.6% | 84.2% |

**三、变量**

本文共有13个变量，变量的定义与类型如下表2所示：

表2

|  |  |  |
| --- | --- | --- |
| **变量名** | **变量描述** | **变量类型** |
| id | 客户号 | integer |
| **bad** | **申请后2年内是否出现过90天以上逾期** | **Y/N** |
| age | 年龄 | integer |
| gender | 性别 | categorical |
| income | 月收入 | real |
| dependent | 家人数量 | integer |
| debtratio | 债务比 | percentage |
| utilization | 信用卡余额所占比例 | percentage |
| num\_line | 贷款和信用卡总数 | integer |
| num\_loan | 房贷笔数 | integer |
| del\_30 | 过去两年逾期30-59天的次数 | integer |
| del\_60 | 过去两年逾期60-89天的次数 | integer |
| del\_90 | 过去两年逾期90天以上的次数 | integer |

本文将bad作为因变量，其余12个变量为自变量。“坏”客户定义为在表现窗口中出现过90天及以上逾期的客户。用于建模的数据共有90101个样本，其中有6075名客户申请后两年内出现过90天以上的逾期，84026名客户未出现过。经过异常值处理、均值插补等方法处理后的数据集共有89590个样本，其中有6042名客户申请后两年内出现过90天以上的逾期，83548名客户未出现过。用于验证的样本共有29971条数据，其中有1998名客户申请后两年内出现过90天以上的逾期，27973名客户未出现过。

**四、建模方法**

建立逻辑回归，b,o,p目前待定。

**五、时间安排**

|  |  |
| --- | --- |
| 4.6 | 小组讨论分工安排。同时进行预先分析，定下样本范围、观测点选择、观察窗口、表现窗口、因变量定义、验证样本以及可利用数据。同时开始写项目计划书。 |
| 4.7 | 进行数据处理，完成异常值处理、遗失值处理、特殊编码、变量推导和变量初选。 |
| 4.8-4.9 | 建立模型，确定变量细分区和变量粗分区。 |
| 4.10 | 完善项目计划书 |
| 4.12-4.14 | 确定评价指标和变量选择 |
| 4.15-4.18 | 模型表现验证和样本稳定验证 |
| 4.19-4.21 | 撰写建模文档 |

**六、项目组成员**

|  |  |
| --- | --- |
| 王婷、迟景升、张杰婷 | 数据预处理；建立模型，确定变量细分区和变量粗分区；  模型表现验证和样本稳定验证； |
| 武泽文、杨嘉敏 | 撰写项目计划书和建模文档，制作汇报PPT |

**七、附件**