**选题2: 互联网金融时代下信用评分体系模型的构建**

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。

征信机构利用采集到的丰富信息对个人进行综合信用评价。在丰富海量的个人信用历史和信用行为数据基础上，采用数据挖掘方法得出的信用行为模式能够更加准确地预测个人未来的信用表现，能够提高操作的效率，降低授信成本，精确估计消费信贷的风险，是金融机构内部评分不可替代的重要工具。因此，建立精准的信用评分体系对于企业有着重要的意义。

对于本题，报名期结束后，组委会将统一发放某贷款机构的历史业务数据作为原始数据，参赛者需要基于给定数据，运用数据挖掘等方法，构造模型变量，制定信用规则，建立信用评估模型，预测违约情况，并最终提交PDF版本的初赛论文正文及数据分析程序。

[1]白云峰,毕强.美国个人信用评分体系研究及启示[J].现代管理科学,2010(12):31-32+38.

美国的FICO信用评分是一个信用分统计模型。FICO信用评分模型中所关注的关键因素主要有五类：包括客户的信用偿还历史、信用账户数、使用信用的年限、正在使用的信用类型、新开立的信用账户。

1.偿还历史。影响FICO得分的首要因素是客户的信用偿还历史。客户的历史偿还情况主要包括各种账户（信用卡、零售商账户、分期付款、财务公司账户及抵押贷款）的支付信息；负面公共记录以及诸如破产、抵押、诉讼、留置等报告事项，账户及应付款的违约情况以及公共记录的细节；支付账户未出现延期的天数等。不良记录发生的时间越近，对总分的影响就越大。

2. 信用账户数。包括仍需要偿还的信用账户总数、信用账户的余额等。

3. 建立信用的年限。一般来讲，客户使用信用的历史越长，越能增加FICO信用得分。因为以往的结算历史记录越多，对未来偿付能力的预测就越准确。

4.新开立的信用账户。历史数据表明，如果一个人经常开立新的信用账户，而不是长期使用固定的信用账户，表明他遵守每个账户信用约定的意愿越低，呈现出更加不稳定的信用水平。

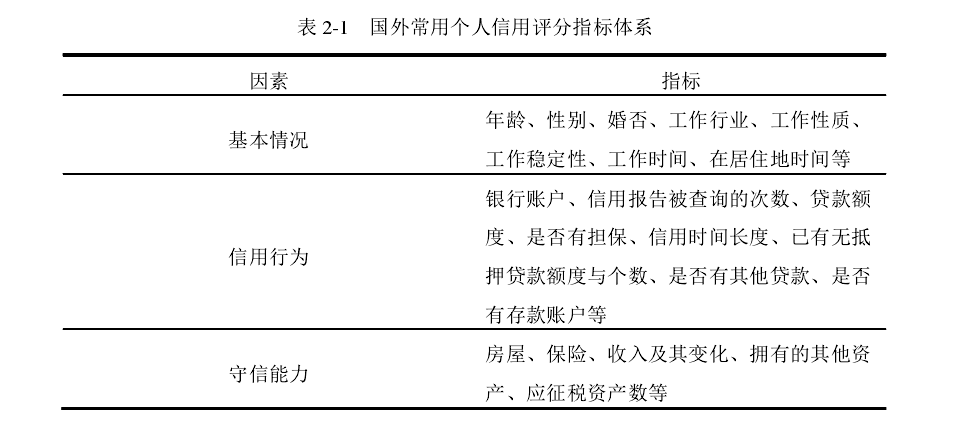
5.拥有的信用类型。该项因素考查客户是否具有一个“健康”的信用组合。包括客户的信用卡账户、零售账户、分期付款账户、金融公司账户和抵押贷款账户的混合使用情况。

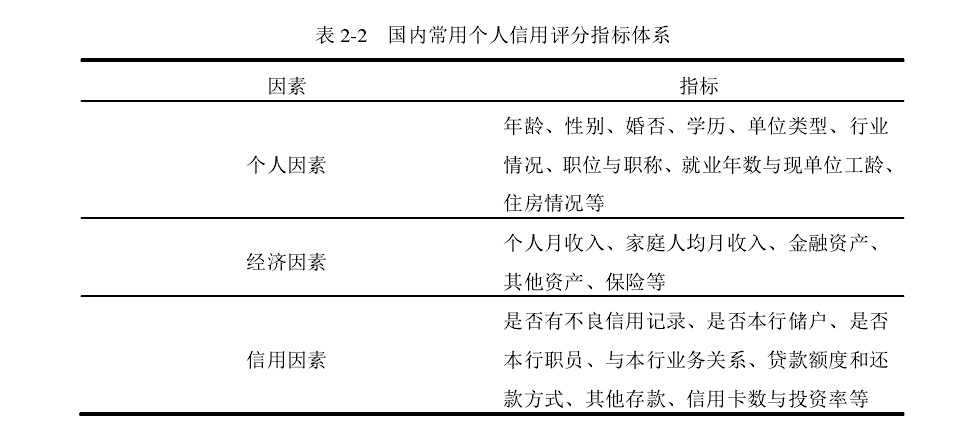
[1]杜淼淼.美国个人信用评分系统及其启示[J].南方金融,2008(08):63-66.

[1]刘焕成,温芳芳.美国个人信用信息管理对我国的启示[J].图书情报知识,2008(06):96-99.

[1]向晖. 个人信用评分组合模型研究与应用[D].湖南大学,2011.

[1]张飞. 个人信用评分组合模型研究[D].电子科技大学,2015.





缺失值预处理等。对于数据缺失问题主要有以下三种处理方法：一是把缺失数据删除；二是将缺失数据用某种方法填充代替；三是将缺失数据项当做一个特征项。比如年龄项缺失时，我们可以将申请人年龄项取为“缺失”项，这样做因为缺失数据包含了某种信息。本文样本数据只有少量缺失，采取均值方法补齐。

[1]吕杨. 个人信用评价体系构建研究[D].南京理工大学,2008.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 变量名称 | 变量类型 | 变量的处理与赋值方式 | 变量对应数据表 | 备注 |
| **个人基本情况** | | | | |
| 是否本地籍 | 虚拟变量 | 0=非本地常住户口,1=本地常住户口 | 基础表 |  |
| 工作城市 | 分类变量 | 共284个城市 |  |
| 教育 | 顺序型变量 | 本科、博士研究生、初中、高中、其他、硕士及以上、硕士研究生、专科、专科及以下 | 分类有点乱，考虑分类合并 |
| 婚姻 | 分类变量 | 离婚、离异、其他、丧偶、未婚、已婚 |  |
| **个人经济状况** | | | | |
| 是否有公积金 | 虚拟变量 | 0=无， 1=有 | 基础表  基础表 |  |
| 收入 | 连续型变量 |  | 缺失严重，缺失约70% |
| **贷款情况** | | | | |
| 客户渠道 | 分类变量 | 共18种 | 基础表 | 缺失严重，缺失约70% |
| 账户状态 | 分类变量 | 呆账、结清、逾期、正常、转出 | 机构版征信-贷款 |  |
| 贷款种类细分 | 分类变量 | 9类贷款 | 机构版征信-信用提示表中统计了个人住房贷款笔数、个人商用房贷款笔数、其他贷款笔数 |
| 担保方式 | 分类变量 | 8种 |  |
| 还款频率 | 数值型变量 |  |  |
| 还款期数 | 数值型变量 |  |  |
| 五级分类 | 顺序型变量 | 次级、关注、可疑、未知、正常 | 缺失0.5305267 |
| 合同金额 | 数值型变量 |  |  |
| 本金余额 | 数值型变量 |  |  |
| 剩余还款期数 | 数值型变量 |  |  |
| 本月应还款 | 数值型变量 |  | 合并取差值 |
| 本月实还款 | 数值型变量 |  |
| 当前逾期期数 | 数值型变量 |  |  |
| 当前逾期金额 | 数值型变量 |  |  |
| 最近24个月还款状态 |  |  | 机构版征信-贷记卡 | 数据为文本符号。信息未知 |
| 个人住房贷款笔数 | 数值型变量 |  | 机构版征信-信用提示 | 与贷款种类细分变量关联 |
| 个人商用房（包括商住两用）贷款笔数 | 数值型变量 |  | 与贷款种类细分变量关联 |
| 其他贷款笔数 | 数值型变量 |  | 与贷款种类细分变量关联 |
| 贷记卡账户数 | 数值型变量 | 取值范围为1至3669 | 特征含义不明白，是否需要计算两种账户数比值作为新的特征 |
| 准贷记卡账户数 | 数值型变量 | 取值范围为0至15 |  |
| 贷款种类 | 分类数据 | 未结清贷款信息汇总、未销户贷记卡信息汇总、未销户准贷记卡信息汇总 | 机构版征信-未销户贷记卡或者未结清贷款 |  |
| 贷款法人数 | 数值型变量 |  |  |
| 贷款机构数 | 数值型变量 |  |  |
| 贷款账户数 | 数值型变量 |  |  |
| 合同金额 | 数值型变量 |  |  |
| 平均单个贷款机构最大合同金额 | 数值型变量 |  |  |
| 平均单个贷款机构最小合同金额 | 数值型变量 |  |  |
| 贷款余额 | 数值型变量 |  |  |
| 已用额度 | 数值型变量 |  |  |
| 最近6个月平均使用额度 | 数值型变量 |  |  |
| **信用情况** | | | | |
| 贷款种类 | 数值型变量 |  | 机构版征信-逾期(透支)信息汇总 |  |
| 贷款逾期笔数 | 数值型变量 |  |  |
| 贷款单月最高逾期总额 | 数值型变量 |  |  |
| 最大贷款时长 | 数值型变量 |  |  |
| 逾期/透支月数 | 数值型变量 |  |  |
| 逾期/透支金额 | 数值型变量 |  |  |
| 查询类别 | 分类变量 |  | 机构版征信-查询记录汇总 | 只有95各report\_id的数据  只有95各report\_id的数据 |
| 查询次数 | 数值型变量 |  |
| 查询机构 | 分类变量 | 4种 | 机构版征信-报告主表 |  |
| 查询原因 | 分类变量 | 3种 |  |

其他问题：时间指标涉及的特征不会分析。下表给出了竞赛数据各表名称以及对应的记录数、涉及的report\_id数，以及各表涉及的时间指标。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 表格名称 | 记录数 | 涉及report\_id数 | 涉及的时间指标 | 备注 |
| contest\_basic\_train  （基础表-训练集）  contest\_basic\_test  （基础表-测试集） | 测试集10000条  训练集30000条 | 共40000 | 放款时间 |  |
| contest\_ext\_crd\_hd\_report  (机构版征信-报告主表) | 40000 | 40000 | 报告生成时间 |  |
| contest\_ext\_crd\_cd\_ln  （机构版征信-贷款） | 357196 | 35594 | 最近一次还款日期  应还款日  发放日期  到期日期 | 不明白这两个数据表的关系，两表变量基本相同，是否利用其中一表的信息即可。另外贷记卡表中币种变量有10种 |
| contest\_ext\_crd\_cd\_lnd  （机构版征信-贷记卡） | 324229 | 39850 | 发放日期  应还款日  最近一次还款日期 |
| contest\_ext\_crd\_is\_creditcue  （机构版征信-信用提示） | 39970 | 39970 | 首笔贷款发放月份  首张贷记卡发卡月份  首张准贷记卡发卡月份 |  |
| contest\_ext\_crd\_is\_sharedebt  （机构版征信-未销户贷记卡或者未结清贷款） | 76246 | 39948 |  |  |
| contest\_ext\_crd\_is\_ovdsummary  （机构版征信-逾期(透支)信息汇总） | 76212 | 25404 |  |  |
| contest\_ext\_crd\_qr\_recordsmr  （机构版征信-查询记录汇总） | 760 | 95 |  | 记录数量过少，是否不考虑该表信息 |
| contest\_ext\_crd\_qr\_recorddtlinfo  （机构版征信-信贷审批查询记录明细） | 654329 | 39865 |  |  |
| contest\_ext\_crd\_cd\_ln\_spl  （机构版征信-贷款特殊交易） | 67725 | 14190 |  |  |
| contest\_ext\_crd\_cd\_lnd\_ovd  （机构版征信-贷记卡逾期/透支记录） | 199644 | 16429 |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| 表格名称 | 记录数 | 涉及report\_id数 | 数据缺失情况（%） |
| contest\_basic\_train  （基础表-训练集）  contest\_basic\_test  （基础表-测试集） | 测试集10000条  训练集30000条 | 共40000 | 训练集0.7198  测试集0.5897 |
| contest\_ext\_crd\_hd\_report  (机构版征信-报告主表) | 40000 | 40000 | 0 |
| contest\_ext\_crd\_cd\_ln  （机构版征信-贷款） | 357196 | 35594 | 0.001220618 |
| contest\_ext\_crd\_cd\_lnd  （机构版征信-贷记卡） | 324229 | 39850 | 0.0009530301 |
| contest\_ext\_crd\_is\_creditcue  （机构版征信-信用提示） | 39970 | 39970 | 月份日期有缺失，对于日期数据间关系不清楚，故并未处理和统计相应的缺失 |
| contest\_ext\_crd\_is\_sharedebt  （机构版征信-未销户贷记卡或者未结清贷款） | 76246 | 39948 | 1 |
| contest\_ext\_crd\_is\_ovdsummary  （机构版征信-逾期(透支)信息汇总） | 76212 | 25404 | 0 |
| contest\_ext\_crd\_qr\_recordsmr  （机构版征信-查询记录汇总） | 760 | 95 |  |
| contest\_ext\_crd\_qr\_recorddtlinfo  （机构版征信-信贷审批查询记录明细） | 654329 | 39865 |  |
| contest\_ext\_crd\_cd\_ln\_spl  （机构版征信-贷款特殊交易） | 67725 | 14190 |  |
| contest\_ext\_crd\_cd\_lnd\_ovd  （机构版征信-贷记卡逾期/透支记录） | 199644 | 16429 |  |

缺失值处理1（单一填补）：<https://wenku.baidu.com/view/4bccef0c31126edb6e1a1007.html>

缺失值处理2（全）：

<https://wenku.baidu.com/view/f6a717c35a8102d277a22fb7.html?mark_pay_doc=0&mark_rec_page=1&mark_rec_position=5&clear_uda_param=1>

<http://www.xueqing.tv/cms/article/185>

缺失值可视化：

http://mp.weixin.qq.com/s/BllrytaoM3RzRRK-V96OZg