**厦门国际银行**

线下零售客户评分卡模型报告

**目 录**

[第一章 导言 5](#_Toc56371278)

[1.1 背景和必要性 5](#_Toc56371279)

[1.2 模型开发目标 5](#_Toc56371280)

[1.3 项目概述 5](#_Toc56371281)

[1.3.1 模型开发步骤 5](#_Toc56371286)

[1.3.2 模型开发 7](#_Toc56371287)

[1.3.3 模型验证 7](#_Toc56371288)

[第二章 数据来源与探索 9](#_Toc56371289)

[2.1 数据来源 9](#_Toc56371291)

[2.2 数据探索 9](#_Toc56371292)

[2.3 排除规则 11](#_Toc56371293)

[2.4 好坏定义 11](#_Toc56371294)

[2.5 窗口定义 14](#_Toc56371295)

[第三章 特征加工 16](#_Toc56371296)

[3.1 候选特征衍生 16](#_Toc56371298)

[3.2 特征分析 16](#_Toc56371299)

[（一） Iv值分析与分箱调整 16](#_Toc56371300)

[（二） 稳定性 16](#_Toc56371301)

[（三） 相关性分析 16](#_Toc56371302)

[（四） 回归显著性分析 16](#_Toc56371303)

[（五） 特征分布 17](#_Toc56371304)

[第四章 模型开发和结果 18](#_Toc56371305)

[4.1 抽样设计 18](#_Toc56371307)

[4.2 模型开发和结果 18](#_Toc56371308)

[4.2.1 特征筛选 18](#_Toc56371309)

[4.2.5模型结果 19](#_Toc56371310)

[4.3 模型的评分设计 22](#_Toc56371311)

[4.4 模型的应用 23](#_Toc56371312)

[附件一 特征变量情况 23](#_Toc56371313)

[附件二 模型开发重要指标及说明 23](#_Toc56371314)

[附件三 逻辑回归(Logistic Regression)概述 26](#_Toc56371315)

**文档版本**

|  |  |  |  |
| --- | --- | --- | --- |
| 日期 | 版本号 | 修改细节 | 作者 |
| 2020/11/14 | 1.0 |  | 风险评估部  数字金融开发中心 |
| 2020/11/20 | 1.0.1 | 3.2特征分析部分增加VIF分析；  3.2附件更新；模型代码增加对逐步回归前的特征分箱进行调整，主要涉及授信总额、平均额度等；  4.2模型开发结果图表更新；  4.4评分数据更新； | 风险评估部  数字金融开发中心 |
| 2020/11/23 | 1.0.2 | 3.2附件更新；  模型代码增加对逐步回归前的特征分箱进行调整，主要涉及查询机构数-担保审查、查询机构数-贷款审批等，同时在逐步回归筛选完变量后对系数偏差的变量进行删除，并再次进行逐步回归，多次进行该项操作，最终保证所有入模变量与业务逻辑一致，同时具有显著性；  4.2模型开发结果图表更新；  4.4评分数据更新；  重新排版； | 风险评估部  数字金融开发中心 |

# 导言

## 背景和必要性

随着零售线下业务的发展和当前经济环境的变化，业务风险特点发生最大的变化，之前的基于专家经验的专家评分卡模型已经不能有效管控线下业务风险。急需引进机器学习等人工智能方法，从大量客户和业务的数据中寻找业务的潜在风险因子，提高零售线下风险管理的精准度和有效性，提高线下零售业务的资产质量。

零售线下业务的概览介绍：

为了有效控制零售线下业务的风险，本项目拟利用逻辑回归方法，通过授信申请信息、个人信息、配偶信息、企业主体信息等信息，进行整合衍生成客户多维度特征，构建线下零售客户申请评分模型，实现人工智能替代专家经验，提高风险管理精准度和有效性。

## 模型开发目标

此次模型开发的目标旨在授信申请信息、个人信息、配偶信息、企业主体信息等信息，通过逻辑回归方法为零售线下客户信用评分，其结果可以应用在零售线下经营贷和消费贷业务场景中，提高行内自动审批的效果，提高审批效率。

## 项目概述

本次零售线下客户申请评分模型使用定义明确的建模方法、流程及验证/考核指标，以保证所建立的模型在数据支持的前提下达到一定的标准，满足项目设计的要求。



### 1.3.1模型开发步骤

本次项目将采用逻辑回归方法进行模型开发，评分模型开发的流程主要包括以下几个步骤：

#### 1.3.1.1 数据分析

对具体建模应用场景下的产品客群分布、数据量、样本时间范围、违约变化情况等进行深入分析。

#### 1.3.1.2 模型设计

对本项目涉及的应用场景，使用业界普遍的建模方法论，分析数据和模型参数，根据分析结果给出模型设计建议，并与相关部门讨论优化，得到项目设计文档，具体步骤如下：

* 根据业界普遍使用的建模方法论，结合行内目前可用数据，依次对模型开发的重要参数进行探索性分析与设计。模型参数包括：排除标准、好/坏/不确定定义、观
* 观测窗口、表现窗口。开发人员将根据数据分析结果提
* 项目设计文档：本文档详细描述了申请模型开发的方法。包括：

1) 建模流程和方法

2) 排除标准

3) 好/坏/不确定定义

4) 观测窗口/表现窗口

5) 模型细分

#### 1.3.1.3 建模数据准备

数据准备的目的是准备模型开发所需的所有数据，该过程需要进行大量的数据清洗和转换工作，包括：

* 根据确定的设计参数，针对模型开发进行数据提取，包括删除无关的记录和变量；
* 对数据进行清洗，清洗的依据主要是逻辑判断、业务经验和相关业务规则，同时对一些不符合规范的数据进行分析处理；
* 根据确定的设计参数，生成目标变量和候选预测变量。

#### 1.3.1.4 建模单元的选择

建模单元的选择，取决于评分卡的应用层级为产品级还是账户级。在决定建模单元时，通常需要考虑以下几方面的因素：

* 风险管理策略的应用层级；
* 申请评分在业务上将如何应用；
* 实施的可行性及复杂性。

#### 1.3.1.5 变量选择

申请评分模型的候选变量数以百计，为提高模型开发的效率，通常会对候选变量进行预筛选。业界最佳实践在进行预筛选时通常考虑变量的稳定性（Population Stability Index，PSI）和区分力[[1]](#footnote-1)（Information Value，IV），此次项目还会结合决策树模型的特征重要性，目的是为发现一些IV和PSI等单变量分析无法发现的变量组合。

变量选择是一个反复研究自变量和因变量之间的一一对应关系、选择最有助于区分好坏的变量的过程。但该过程并不是一个单纯的统计决策的过程。业务知识也在很大程度上有助于变量集的选择。

### 1.3.2 模型开发

细分方案确定后，开发人员将针对每个细分人群分别构建样本数据、设计衍生变量、开发评分模型，在模型开发的过程中需要注意以下事项：

* 为了减小缺失值和异常值对建模的影响，通常在建模前对连续型变量进行处理；
* 为了更好地拟合自变量与因变量之间的关系，增加模型的预测力，可对类别型变量产生WOE，对连续型变量进行函数变形，并使用与因变量关系最紧密的函数形态作最终拟合；
* 最终模型变量的选择并不是一个单纯的统计决策，业务认知和对行内数据的了解将在很大程度上影响最终模型的形态；
* 使用逻辑回归方法，进行模型开发，并确认最终进入模型的变量和系数。

本次申请评分优化项目将采用 KS/AUC进行模型区分力的评估，具体详见附件二。

### 1.3.3 模型验证

本次申请评分优化项目将根据数据的可行性进行两项验证，确保模型的稳健性：

#### 1.3.3.1 预留样本验证

本次申请评分优化项目将预留样本验证作为评分模型开发过程的一部分。在建模时点随机选取80%的样本作为评分模型的开发样本，并将模型结果应用到其余20%的预留样本上，以检验模型是否稳定有效。

在预留样本上验证模型的目的，是通过未被用于建模过程的独立样本来判断评分模型的预测力。如果模型在预留样本和开发样本上的预测力存在较大差异，说明开发过程中存在过度拟合(Over-Fit)，该模型在实际应用环境中将不能很好地区分好坏。

#### 1.3.3.2 跨时间样本验证（OOT）

在数据支持的情况下，本次申请评分优化项目将对评分模型进行跨时间样本验证。跨时间样本验证将所开发的模型应用到不同时点的样本上，以检验模型是否稳定有效。

需要注意的是，人群的变化会在一定程度上会影响跨时间样本验证的效果。

#### 1.3.3.3 评分模型的校准

评分校准是将评分模型的结果映射到特定好坏比率的过程。为了方便日常的业务管理，一般通过转化，在评分结果与特定的风险水平之间建立起一定的函数关系（如：在评分为 400 时，好坏比率为 1:1；每增加 40分，好坏比率加倍）。本次申请评分优化项目将使用如下的评分校准方式：在评分为 600 时，好坏比率为 20:1；每增加 20分，好坏比率加倍。

|  |  |
| --- | --- |
| 分数 | 好坏比 |
| 600 | 20：1 |
| 620 | 40：1 |

#### 1.3.3.4 模型文档准备

为便于模型的正确实施与存档，最终模型开发的结果、变量的来源及算法将反映在模型文档中。

# 第二章 数据来源与探索



## 数据来源

数据来源主要来自我行信贷系统录入的授信申请信息、个人客户信息、配偶信息、，模型开发通过脱敏数据调阅的方式，由科技开发部提供脚本，科技运维部协助从生产中获取。

## 标签定义

### 2.4.1 设计方法

统计模型的建立是为了预测预定义的某种结果。典型的风险模型可以“告知”以下一种或几种事件发生的可能性：破产损失、信用损失、严重逾期。重要的是将信用周期内不同阶段风险管理的业务需求转换成可以计量的模型目标。好坏定义必须兼顾数据分析的结果和业务管理的需要。另外，为了增强模型对好坏的区分能力，通常在定义好坏客户/账户的同时，将介于好坏状态之间的部分定义为不确定，从模型拟合样本中剔除。

本次项目中，在定义好坏时需要考虑以下几方面的因素：

* 巴塞尔新资本协议中零售违约定义：逾期90天以上、核销等；
* 银行的业务策略：催收策略、核销政策等；
* 滚动率分析；
* “坏”群体的数量：最终定义为“坏”的群体数量不能过少；
* 数据对于相关定义的支持度：最终好坏定义必须能在现有数据上实施。

通常情况下，当个人贷款产品达到逾期90天以上的状态时，借款人财务状况恶化、支付出现困难，且难以获得新的资金。同时，为减少误判，在好坏分析过程中，会参考滚动率分析的结果对处于好坏边缘的客户做出判断，并且将某些难以界定的个体归为不确定。滚动率分析是通过考察不同类别间群体的滚动情况，辅助好坏边缘客户的判断，帮助确定合适的好坏定义。

### 2.4.2 滚动率分析

好坏定义应该具有一定的稳定性，即在某一时间段内被定义为“坏”的人群在未来也表现为“坏”，反之亦然。为了判断好坏定义的稳定性，通常采用滚动率分析对好坏定义的恰当性进行评估。滚动率分析是分析贷款产品的逾期状态在不同时间段的滚动情况, 从而得知相邻时间段账户往前滚动（恶化）、往后滚动（改善）以及保持原状的数量，可用于观察账户的滚动情况是否稳定。其核心内容是通过产生、分析同一群体在两个相邻时间窗口好坏表现的转移矩阵，考察不同类别间的滚动情况，并根据滚动的趋势及比例，帮助调整并确定合适的好坏定义。分析过程可归纳为：

取两个相邻的时间窗口，在第一个时间窗口中对账户根据逾期期数以及次数进行分类，划分为不同群体（如无逾期，逾期1期1次，逾期1期2次以上等），并统计第一时间窗口中各群体在第二个时间窗口内的滚动变化情况。若第一时间窗口内某一类别群体绝大部分会在第二个窗口中落入更“坏”的类别，则将该类别定义为“坏”；反之，若第一时间窗口中某一类别群体中绝大部分会在第二个窗口落入更“好”的类别，则将该类别定义为“好”。其他则定义为“不确定”。以下为本次项目滚动率分析的过程：

为了更加准确地判断好坏的定义，取2020年2月29间有余额的客户，分析其从当月到下个月的逾期滚动率情况，分析结果如下：

因为该产品有保险赔付，在逾期超过六十天后后就会出现保险赔付，所以我们认定有出现保险理赔的客户也属于损失客户。

**表 1 授信申请信息滚动率分析**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2020.7 | | | | | | | | | |
| 2020.6 | M0 | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8+ | 总计 |
| M0 | 99.06% | 0.93% | 0.01% |  |  |  |  |  |  | 100.00% |
| M1 | 35.29% | 33.33% | 29.41% | 1.96% |  |  |  |  |  | 100.00% |
| M2 | 13.89% | 8.33% | 11.11% | 66.67% | 0.00% |  |  |  |  | 100.00% |
| M3 | 15.56% | 0.00% | 2.22% | 35.56% | 46.67% | 0.00% |  |  |  | 100.00% |
| M4 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 100.00% | 0.00% |  |  | 100.00% |
| M5 | 33.33% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 50.00% | 16.67% |  | 100.00% |
| M6 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 100.00% | 0.00% | 100.00% |
| M7+ | 5.26% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 94.74% | 100.00% |
| 总计 | 97.09% | 1.19% | 0.30% | 0.59% | 0.30% | 0.10% | 0.04% | 0.13% | 0.26% | 100.00% |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 更好 | 保持现状 | 更坏 |
| M0 |  | 99.06% | 0.94% |
| M1 | 35.29% | 33.33% | 31.37% |
| M2 | 22.22% | 11.11% | 66.67% |
| M3 | 17.78% | 35.56% | 46.67% |
| M4 | 0% | 0% | 100% |
| M5 | 33.33% | 0% | 66.67% |
| M6 |  | 0% |  |
| M7+ | 5.26% | 0% | 94.74% |

由上表数据分析结果可以看出来：

* 上期从未逾期的客户约99.06%在下一期观察点中仍表现为未逾期；
* 上期仅有过一次M1逾期的客户仅有35.29%在下一期观察点口中会变为未逾期，但也有31.27%在下一期观察窗口中会变为M2+;
* 上期逾期M2的客户在下一期观察窗口逾期状态变得更坏的比例为66.67%；
* M3的变化有35.56%保证一致，经过分析，这些客户M3已经维持了两个月以上，风险情况已经很差，大部分已无力偿还，由于分行催收和不良的压力，导致一直维持在M3。

从上述滚动趋势来看，从未逾期的账户更倾向于在未来变得更好而不是更坏，从逾期状态为M1开始的账户变成坏的概率只在50%左右，有很大不确定性，而从逾期状态为M2的客户倾向在未来变得更坏而不是更好。参考巴塞尔新资本协议和行内现有产品的好坏定义，并结合业内比较通用的定义。我们建议：将表现期从未逾期的账户定义为好账户，表现期最大逾期状态为M2的账户定义为坏账户。

**2.4.2.3 设计结果**

根据滚动率分析结果，我们将坏客户定义为逾期状态M2及以上，其他定义为好客户。

## 窗口定义

观察窗口：观察窗口是模型开发过程中选定的样本进件时间段，在该窗口内的进件将作为模型开发和样本内验证的客群基础。观察窗口的选取通常需要考虑产品自身发展的特点以及季节性效应等对产品申请的影响。

表现窗口：在既定的观察窗口内选定的建模群体的表现，即好/坏/不确定的界定，必须在一个设定的时间段内定义，这一时间段即申请评分模型的表现窗口。窗口的长度可随不同的产品类别、模型类别、数据可获取性而不同，并在很大程度上取决于业务目标。

一般来说，表现窗口的选取应足够长，使得绝大多数的坏账户有机会将逾期行为表现出来。然而表现窗口越长，就意味着观察窗口的临近性越差。因此即使有很长周期的历史数据，在确定窗口时也需要考虑完整性（能捕捉多少坏账户）和及时性（与当前业务策略的相关性）之间的平衡。

在选择表现窗口时需要考虑以下几方面的因素：

* 用于模型开发样本的时间临近程度及与未来应用的相关性；
* 选定观察点样本数量足以在表现窗口中产生足够的好和坏个体；
* 数据源的限制，即需要保证有充足、可靠的预测变量；
* 产品期限特征；
* 账龄分析。

结合线下零售产品的实际情况，线下零售产品从2017年就开始上线，但由于前期征信的数据保存的问题，本次建模从2018年10月开始抽取申请，并抽取表现到2020年9月的表现数据。



可以看出平均余额M2+到MOB14开始趋于稳定，为了提前识别客户风险，基于专家经验，本次表现期定义成放款后13个月。

## 数据探索

1.数据探索包括数据存储情况、原始数据检测、表格关联关系与交叠分析等，为数据预处理与特征工程提供数据分析基础。数据探索详见《数据质量分析报告》，如附件二。

2.根据数据质量分析情况，线下零售贷款数据整体数据质量较差，相关数据表匹配率仅在40%，关联程度一般。表内字段存在缺失率较高、具有单一值完全相同等情况，将在后续数据预处理中进行调整。

## 抽样设计

### 2.5.1 排除规则

年审未审查的客户

无提款支用的客户

无历史征信的客户

### 2.5.2 抽样设计

基于前期对数据的分析结果，线下零售客户在2017开始申请客户，但由于行里保存的征信数据只能保存两年，征信数据只保留到2018年10月，同时2020年1月之后开始使用二代征信，且表现期定义为13个月，因此本次建模选择2018年10月至2020年1月申请的客群为样本。本次对样本采用不抽样处理，对全量样本分训练集、测试集和时间外验证集三块来建模。

具体样本区分：

1. 使用的总训练样本数量为6154条，坏样本275条，好样本数为5879条，，坏样本率为4.7147%
2. 本次建模以2018年10月-2019年8月的建模样本为训练和测试集，以2019年9月-2020年1月为时间外样本。其中训练集和测试集以8:2的比例随机抽取。

# 第三章 特征工程



## 候选特征衍生

线下零售评分卡模型衍生特征主要是通过对申请信息的二次交叉衍生，征信特征的累计加总，地址信息编码聚类，类别特征编码，变量分箱等方法来进行处理。

## 特征分析

### Iv值分析与分箱调整

Iv值衡量了特征区分好坏客户的能力，对特征分箱后计算每个特征的iv值以及查看woe的单调性，首先根据iv值筛选出iv值大于0.01的特征，然后再根据业务逻辑筛选woe呈现单调递增递减或U型、倒U型的特征，同时对IV值较高，但是不满足单调性的特征进行分箱调整使其满足业务逻辑要求。（见附件四《特征IV值分析》）

### 稳定性

变量的不稳定将影响样本客群的代表性和评分的稳定性，最终导致模型整体的不稳定。变量稳定性以群体稳定指数（PSI）衡量，通常通过计算变量在开发样本和跨时间样本上的群体稳定性指标对变量的稳定性进行评估。

线下零售评分卡模型特征PSI指数是以2018年10月至2019年9月为基准期， 2019年10月至2020年1月为观测月进行计算。根据PSI指数，将选取稳定性较好（均小于0.1）的特征作为入模特征。（见附件五《特征PSI分析》）

### 相关性分析

线下零售评分卡模型是基于逻辑回归模型进行开发的，因此特征间的共线性对模型具有很大影响，因此，在特征筛选过程中，对模型sperman相关系数大于0.7的特征进行剔除，保留iv值较高的特征。见附件六《特征相关性分析》

### VIF分析

在相关性分析完成后，针对部门相关性较低，但是依旧会在线性模型中产生共线影响的变量采用计算VIF，然后剔除VIF大于5的特征。详见附件七《特征VIF分析》。

### 回归显著性分析

逻辑回归是属于广义线性回归模型的一种，而对于线性回归模型来说，特征的显著性（P值）非常重要，在线性模型中，我们需要每个进入的特征都对目标变量具有显著效果，因此在特征筛选中，采用逐步logit回归的方式，剔除对目标变量显著性（P值）大于0.1的特征。见附件八《特征逐步回归结果》

### 特征分布

特征分布将直观的展现每个特征是否与业务合理性相吻合，基于线下零售评分卡模型入模特征分布见附件九《入模特征分布》。

## 特征筛选

特征筛选主要通过以下几个步骤：

1. 根据业务含义对一些意义不明和加工后原特征可替代的特征进行筛除；
2. 对分箱后IV值低于0.01的特征进行剔除，并对iv值大于0.01但分箱后woe走势不符合业务逻辑的特征箱体进行调整；
3. 对相关性超过0.6的特征剔除，保留其中iv值最大的特征；
4. 对VIF大于5的特征进行剔除
5. 根据逐步回归对回归P值大于0.1的特征进行剔除；
6. 根据PSI对稳定性较差（超过0.1）的特征进行筛除；

特征衍生后共计有180个特征，根据业务含义筛除71个字段（如户籍地与居住地市级是否一致、配偶是否有单位记录、单位所属行业、性别、年龄等），剩余107个字段；再根据iv值剔除22个，剩余字段85个；根据单一值剔除14个，相关性剔除47个，剩余字段24个；根据VIF大于5剔除特征2个；根据逐步回归显著性剔除10个，剩余字段12个，系数调整删除2个后再逐步回归删除2个，最后剩余8个字段。

# 第四章 模型设计与开发



## 模型设计

本评分卡模型应业务部门要求，原设计通过xgboost重要度来筛选变量，同时对特征采用决策树分箱，后发现因线下特征质量较差，xgboost筛选后特征数量较少，后仅采用决策树来进行分箱，评分卡建模采用逻辑回归。

逻辑回归（Logistic Regression）是研究因变量为二分类或多分类观察结果与影响因素（自变量）之间关系的一种多变量分析方法，属概率型非线性回归。相关算法介绍见附件。

## 模型调优

因样本数量较少，模型开发过程中并未进行模型调优工作。

## 模型变量校验

### （一）合理性校验

本次评分卡模型共涉及到8个入模变量，最终系数如下：

**表4.1 入模变量回归系数**

|  |  |
| --- | --- |
| Variable | Coef |
| mr\_贷款历史机构数\_woe | 0.638728 |
| 近2年内查询次数-担保资格审查\_woe | 0.435455 |
| 近24月累计逾期\_woe | 0.540683 |
| found\_until\_apply\_month\_woe | 0.742516 |
| mr\_贷款最大还款月数\_woe | 0.750525 |
| 近2年内查询机构数-贷款审批\_woe | 0.52692 |
| 住房贷款占比\_woe | 0.504553 |
| 贷记透支比例\_woe | 0.36035 |

经与业务部门验证后，所选变量系数符号全部符合业务逻辑。

### （二）相关性和多重共线性

模型入模变量的相关系数和VIF检验结果如下：

**表4.2 入模变量相关性系数**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | mr\_贷款历史机构数\_woe | 近2年内查询次数-担保资格审查\_woe | 近24月累计逾期\_woe | found\_until\_apply\_month\_woe | mr\_贷款最大还款月数\_woe | 近2年内查询机构数-贷款审批\_woe | 住房贷款占比\_woe | 贷记透支比例\_woe |
| mr\_贷款历史机构数\_woe | 1 | 0.031465 | 0.053535 | 0.04102 | 0.021252 | 0.113681 | 0.014901 | 0.103741 |
| 近2年内查询次数-担保资格审查\_woe | 0.031465 | 1 | 0.104756 | 0.192874 | 0.007134 | 0.33671 | 0.090503 | 0.192492 |
| 近24月累计逾期\_woe | 0.053535 | 0.104756 | 1 | 0.160394 | -0.00302 | 0.23883 | 0.027085 | 0.203802 |
| found\_until\_apply\_month\_woe | 0.04102 | 0.192874 | 0.160394 | 1 | 0.014554 | 0.30272 | 0.00361 | 0.183113 |
| mr\_贷款最大还款月数\_woe | 0.021252 | 0.007134 | -0.00302 | 0.014554 | 1 | 0.092707 | 0.017309 | 0.051096 |
| 近2年内查询机构数-贷款审批\_woe | 0.113681 | 0.33671 | 0.23883 | 0.30272 | 0.092707 | 1 | 0.188551 | 0.440613 |
| 住房贷款占比\_woe | 0.014901 | 0.090503 | 0.027085 | 0.00361 | 0.017309 | 0.188551 | 1 | 0.101196 |
| 贷记透支比例\_woe | 0.103741 | 0.192492 | 0.203802 | 0.183113 | 0.051096 | 0.440613 | 0.101196 | 1 |

**表4.3入模变量VIF检验**

|  |  |
| --- | --- |
| Variable | VIF |
| mr\_贷款历史机构数\_woe | 1.017499 |
| 近2年内查询次数-担保资格审查\_woe | 1.144401 |
| 近24月累计逾期\_woe | 1.084526 |
| found\_until\_apply\_month\_woe | 1.128277 |
| mr\_贷款最大还款月数\_woe | 1.01045 |
| 近2年内查询机构数-贷款审批\_woe | 1.499114 |
| 住房贷款占比\_woe | 1.04241 |
| 贷记透支比例\_woe | 1.267718 |

### （三）特征分布

特征badrate分布符合单调性与非W型走势，具体可参见附件二《特征变量分箱》

# 第五章 模型验证

## 模型区分/排序能力指标

本模型在测试集和训练集AUC以及KS指标表现测试情况如下表所示。建立的LR模型在时间内外验证的AUC满足在0.8087，KS在0.5129左右，模型性能指标结果较为优秀，建立的模型对好坏样本以及对样本的排序能力较强，且模型的稳定性与泛化能力满足要求。

1. 模型整体好坏样本分布如下:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | good | bad | total | goodprob | badprob |
| 测试集 | 2474 | 166 | 2640 | 93.71% | 6.29% |
| 训练集 | 619 | 42 | 661 | 93.65% | 6.35% |
| 时间外 | 2786 | 67 | 2853 | 97.65% | 2.35% |

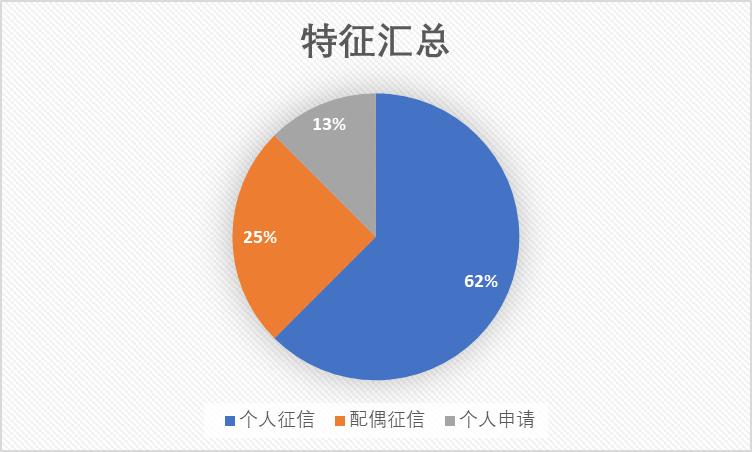
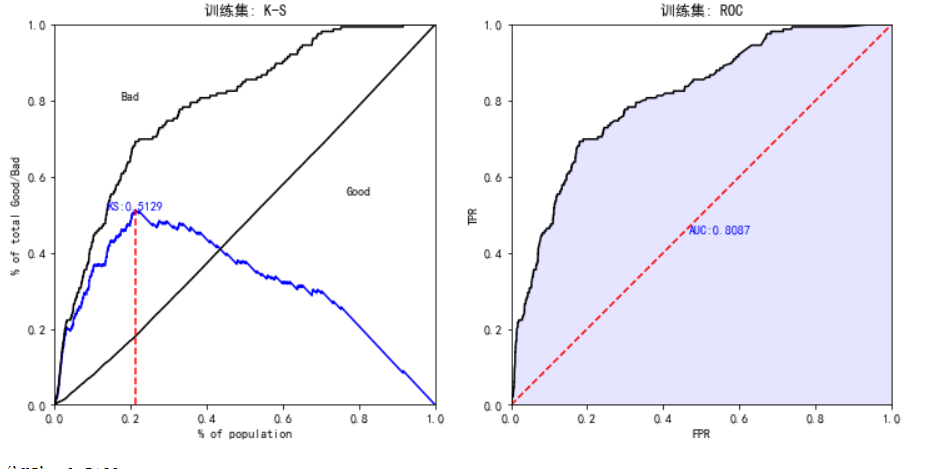
1. 入模8个特征指标，指标分布如下：
2. 评估指标统计结果如下：

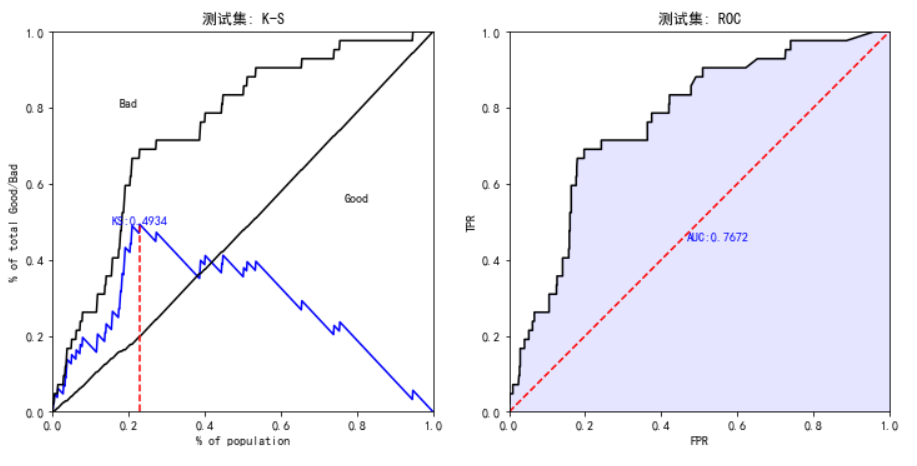
表 6 模型验证的评价指标

|  |  |  |  |
| --- | --- | --- | --- |
| 指标 | 训练集 | 测试集 | 时间外 |
| AUC | 0.8115 | 0.7672 | 0.8394 |
| KS | 0.5129 | 0.4934 | 0.531 |
| PSI |  | 0.0199 | 0.104 |

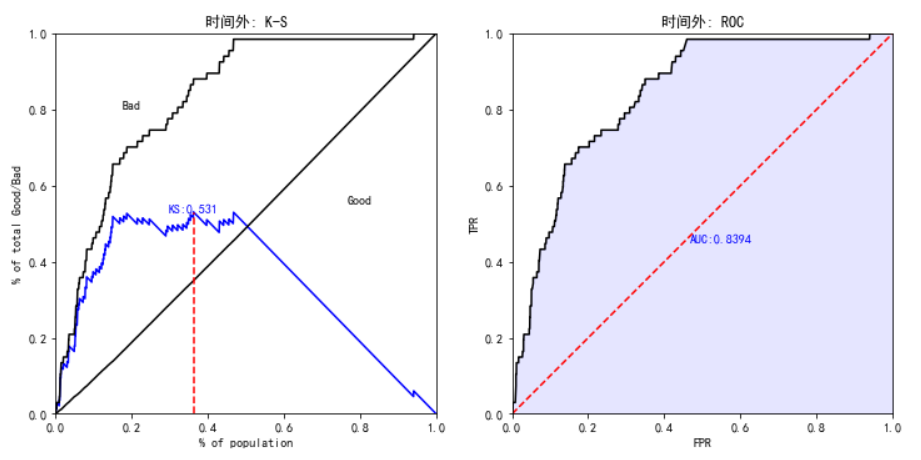
（a）训练集ROC与KS



（b）测试集ROC与KS



（c）时间外ROC与KS

****

## 5.2 模型的稳定性指标

入模特征与模型评分的PSI如下：

1. 测试集：

|  |  |
| --- | --- |
| variable | PSI |
| 贷记透支比例\_points | 0.000486 |
| 近2年内查询次数-担保资格审查\_points | 0.00081 |
| 近24月累计逾期\_points | 0.001804 |
| 近2年内查询机构数-贷款审批\_points | 0.003478 |
| mr\_贷款历史机构数\_points | 0.005945 |
| found\_until\_apply\_month\_points | 0.009521 |
| mr\_贷款最大还款月数\_points | 0.010068 |
| 住房贷款占比\_points | 0.015387 |
| score | 0.019925 |

1. 时间外：

|  |  |
| --- | --- |
| variable | PSI |
| mr\_贷款最大还款月数\_points | 0.002566 |
| mr\_贷款历史机构数\_points | 0.003199 |
| 近2年内查询次数-担保资格审查\_points | 0.005229 |
| 住房贷款占比\_points | 0.015272 |
| 贷记透支比例\_points | 0.016875 |
| 近24月累计逾期\_points | 0.018427 |
| 近2年内查询机构数-贷款审批\_points | 0.024882 |
| score | 0.104441 |
| found\_until\_apply\_month\_points | 0.90038 |

## 5.3 模型的评分分布

参考线上风控模型产品的专家经验，评分卡中为了让评分具有较好的区分能力，应设置坏客户概率/好客户概率=0.05时的基础分为600，当上述比例翻倍时，即坏客户概率/好客户概率=0.1时的分数记录基准为20，因此模型输出违约概率对评分的拟合公式为：

根据评分公式，全量数据集的评分情况如下：

**表 7 全量数据集的分段客户分布情况**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 分数段 | 客户总数 | 坏客户数 | 好客户数 | 坏客户率 | 坏客户数占坏客户总数占比 | 好客户数占客户总数占比 |
| (326.999, 503.3] | 616 | 116 | 500 | 18.83% | 42.18% | 8.50% |
| (503.3, 549.0] | 622 | 74 | 548 | 11.90% | 26.91% | 9.32% |
| (549.0, 585.0] | 633 | 19 | 614 | 3.00% | 6.91% | 10.44% |
| (585.0, 614.0] | 639 | 19 | 620 | 2.97% | 6.91% | 10.55% |
| (614.0, 637.0] | 587 | 17 | 570 | 2.90% | 6.18% | 9.70% |
| (637.0, 662.0] | 603 | 17 | 586 | 2.82% | 6.18% | 9.97% |
| (662.0, 711.0] | 631 | 10 | 621 | 1.58% | 3.64% | 10.56% |
| (711.0, 762.0] | 618 | 0 | 618 | 0.00% | 0.00% | 10.51% |
| (762.0, 789.0] | 827 | 2 | 825 | 0.24% | 0.73% | 14.03% |
| (789.0, 960.0] | 378 | 1 | 377 | 0.26% | 0.36% | 6.41% |

表 8 时间外验证集的分段客户分布情况

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 分数段 | 客户总数 | 坏客户数 | 好客户数 | 坏客户率 | 坏客户占坏客户总数占比 | 好客户数占好客户总数占比 |
| (326.999, 510.0] | 288 | 31 | 257 | 10.76% | 46.27% | 9.22% |
| (510.0, 558.0] | 284 | 16 | 268 | 5.63% | 23.88% | 9.62% |
| (558.0, 597.0] | 303 | 5 | 298 | 1.65% | 7.46% | 10.70% |
| (597.0, 627.0] | 270 | 8 | 262 | 2.96% | 11.94% | 9.40% |
| (627.0, 659.0] | 329 | 6 | 323 | 1.82% | 8.96% | 11.59% |
| (659.0, 694.0] | 243 | 0 | 243 | 0.00% | 0.00% | 8.72% |
| (694.0, 753.4] | 280 | 0 | 280 | 0.00% | 0.00% | 10.05% |
| (753.4, 789.0] | 654 | 0 | 654 | 0.00% | 0.00% | 23.47% |
| (789.0, 960.0] | 202 | 1 | 201 | 0.50% | 1.49% | 7.21% |

## 5.4 模型的应用

将上述模型运用到零下线下个人经营贷系统，对从2018年10月至2020年1月之间所有客户合计进行评分，若将上述模型策略运用在生产实践，对所有客户按逾期可能性从高到低进行排序，从上表可以看出，在503分段左右，坏客户在分段内数量下降，因此设置若直接拒绝分值设置在503和549分，拒绝客户数占比情况如下：

表 9 评分拒绝策略下的逾期和客户数变化  
（基于上述全量训练数据集）

|  |  |  |
| --- | --- | --- |
| 拒绝分数 | 拒绝掉坏客户占比率 | 拒绝掉好客户占比率 |
| 505 | 42.18% | 8.5% |
| 551 | 69.09% | 17.82% |

可以看出来，如果我们把拒绝分数设置在503的话，会抓住42.18%的坏客户，但拒绝掉8.5好客户，可以帮助坏客户率从18.83%下降至11.9%。如果我们把拒绝分数设置在549的话，会抓住69.09%的客户，而拒绝掉17.82%的好客户，帮助坏客户率从18.83%下降到3%。

# 附件一 算法概述



# 附件二 数据质量分析报告



# 附件三 特征变量情况



# 附件四 模型开发重要指标及说明



# 附件五 特征IV分析





# 附件六 特征PSI分析



# 附件七 特征相关性分析



# 附件八 特征VIF分析：



# 附件九 特征逐步回归分析:



# 附件十 入模特征分布



1. PSI 和 IV 的计算公式请参照附件二。 [↑](#footnote-ref-1)