****

**厦门国际银行机器学习**

**通用评分卡模型开发报告**

**目录**

[第一章 导言 5](#_Toc33479373)

[1.1 背景和必要性 5](#_Toc33479374)

[1.2 模型开发目标 5](#_Toc33479375)

[1.3 文档内容 5](#_Toc33479376)

[第二章 数据来源与探索 6](#_Toc33479377)

[2.1 数据来源 6](#_Toc33479378)

[2.2 标签定义 6](#_Toc33479379)

[2.3 窗口定义 6](#_Toc33479380)

[2.4 数据探索 7](#_Toc33479381)

[2.5 抽样设计 7](#_Toc33479382)

[2.6 数据预处理 7](#_Toc33479383)

[第三章 特征工程 8](#_Toc33479384)

[3.1 候选特征衍生 8](#_Toc33479387)

[3.2 特征分析 8](#_Toc33479388)

[（一） 显著性 8](#_Toc33479389)

[（二） 稳定性 8](#_Toc33479390)

[（三） 相关性分析 8](#_Toc33479391)

[（四） 专家回选（或有） 9](#_Toc33479392)

[（五） 特征分布 9](#_Toc33479393)

[（六） *其他（或有）* 9](#_Toc33479394)

[3.3 特征筛选 9](#_Toc33479395)

[3.4 结果 9](#_Toc33479396)

[第四章 模型设计与开发 9](#_Toc33479397)

[4.1 模型设计 9](#_Toc33479400)

[4.2 模型调优 9](#_Toc33479401)

[4.3 模型变量校验（或有，主要针对逻辑回归） 10](#_Toc33479402)

[（一）合理性校验 10](#_Toc33479403)

[（二）相关性和多重共线性 10](#_Toc33479404)

[（三）特征分布 10](#_Toc33479405)

[第五章 模型验证 10](#_Toc33479406)

[5.1 模型验证方法 10](#_Toc33479411)

[5.2 模型区分/排序能力指标 10](#_Toc33479412)

[5.3 模型准确性/审慎性指标（一般用于评级模型、评分分段） 12](#_Toc33479413)

[5.4 模型稳定性指标 13](#_Toc33479414)

[5.5 评分分布矩阵 13](#_Toc33479415)

[5.6 集中度 14](#_Toc33479416)

[第六章 模型优化与迭代（或有） 15](#_Toc33479417)

[附件一 算法概述 15](#_Toc33479418)

[附件二 数据质量分析报告 15](#_Toc33479419)

[附件三 特征工程设计 15](#_Toc33479420)

[附件四 模型评级指标及业界标准 16](#_Toc33479421)

**文档版本**

|  |  |  |  |
| --- | --- | --- | --- |
| 日期 | 版本号 | 修改细节 | 作者 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# 导言

## 背景和必要性

*【说明】该部分主要是对模型开发的背景、项目的意义、使用的方法等进行说明*

*【参考】：*

*由于近年来我行XXXXX业务的不断扩大和风险管理体系的逐步完善，现有XXX分布跟开发样本相比产生一定偏移，可能导致模型无法捕捉客户最新风险特征，出现区分能力下降，客户分布过于集中等现象。从更好助力业务转型发展的角度，有必要进行模型优化升级工作。*

## 模型开发目标

*【说明】该部分主要是对模型开发目标进行说明，是新建模模型还是对原有模型的优化，是否引入新数据或新方法、KS、AUC等模型评估指标预计达到或提升多少，预期对业务开展起到哪些促进作用等等*

# 数据来源与探索

## 数据来源

*【说明】介绍数据来源于哪些内/外部系统，涉及哪些表，哪些字段，字段类型等，可单独形成一份EXCEL嵌入。*

*对源数据字段的来源进行说明，一般包括银行内部经营交易数据、外部第三方数据、合作方推送的数据等，每个数据源包含了多少字段、每个数据字段的类型、数据源的数据质量（可参照外部数据验证规范指引）如何等等。--参考《数据质量分析报告》*

## 标签定义

*【说明】介绍好坏样本是如何进行分析和定义。例如采用逾期滚动率、迁移率等分析方式*

## 窗口定义

*【说明】针对相关模型分为观察窗口和表现窗口，其他场景根据实际情况调整，并提供选择窗口的依据：*

*【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人群基础。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，采用账龄分析，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。需要说明确定表现窗口的分析依据。*

## 数据探索

*【说明】对行内当前的数据进行深入研究和探索，包括单变量、多变量分析，了解数据是否满足要求，并对数据质量进行评估。*

***数据质量分析报告****：此文档基于当前可用的行内数据，说明已完成的数据分析工作及发现，作为附件二。*

## 抽样设计

*【说明】介绍在模型开发时是如何进行抽样设计，使用了哪些方法，训练集、验证集和测试集是如何进行采样的。包括每个集合的样本占比、时间跨度、使用方法等。*

*【参考】*

|  |  |  |  |
| --- | --- | --- | --- |
| *样本类型* | *样本数量和占比* | *样本时间跨度* | *采用方法* |
| *0-可疑样本* | *2000，10%* |  |  |
| *1-正常样本* | *20000，90%* |  |  |
|  |  |  |  |

## 数据预处理

*【说明】针对分析结果对哪些字段进行数据预处理，包括缺失值、异常值是如何处理，连续变量使用什么方法进行分箱处理。*

*例如：业界通常根据变量取值进行分组（如将年龄分成5等分），计算每个分组的好坏比率，并经归并后，建立相应的“WOE变量”，以满足后续的分析需要。在分组归并时通常考虑以下原则：*

* *好坏比率的总体变化趋势与业务经验一致；*
* *好坏比率的总体变化趋势最好单调，线性更佳。如果业务或数据很有把握也可以U Shape，但V Shape最好合并成单调；*
* *分组总数不宜太多（通常5-8组）；*
* *各分箱人群比例不宜太小（<5%或坏样本<50），也不宜过度集中(>25%)，除非有特殊情况；*
* *若WOE差距小于0.1(GBI index小于1.1倍)，建议合并；*
* *用于Segment的变量分界点两侧WOE最好一正一负；*
* *WOE通常在-2到+2之间，若超过为极强变量，使用中要谨慎；*
* *Missing、特殊值和异常值都必须是独立分箱，单独分析原因定性赋予一个WOE值，不建议与有值的客群合并。*

# 特征工程



## 候选特征衍生

*【说明】介绍特征工程中针对哪些候选特征使用了哪些方法进行特征衍生，这里只是进行总体说明，详细的特征衍生说明作为****附件三：《特征设计清单》****提供。*

## 特征分析

*【说明】介绍特征分析过程使用了那些方法，但不限于以下几种。*

### 显著性

在候选变量集中，选取对目标变量区分能力较高的变量子集，区分力用信息值（Information Value，IV）衡量。*给出IV阈值，提供分析过程和筛选结果*。

### 稳定性

在模型中使用不稳定的变量是不可取的，将影响样本人群的代表性和评分的稳定性。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。*给出PSI设定阈值，并提供筛选结果，分析结果可形成独立报告作为附件。*

### 相关性分析

*由于特征的强相关性会影响模型的准确度和稳定性，需要进行各变量的相关性分析。介绍分析方法和分析结果。*

### 专家回选（或有）

*一部分在上述过程中删除的变量，在业务和经验上具有较大的意义。可以根据专家判断将这部分变量回选，通过分组和变换处理，尝试放入模型。提供回选结果。*

### 特征分布

*特征加工完之后，分别分析好、坏样本的特征分布，如果特征是连续型的，主要分析均值、方差、缺失占比、四分点等统计值，如果是离散型，则考察各分类取值的频数占比，提供对哪些特征进行分布分析、特征分布情况应特征设计清单一部分。*

### *其他（或有）*

## 特征筛选

*【说明】介绍使用哪些方法进行特征筛选，作为最终入模的特征变量。*

## 结果

特征工程设计结果详见附件三：《特征工程设计》。*其中应包含3.1章节特征衍生字段说明：包括衍生字段、使用方法、原始字段等；3.2章节特征（包含衍生特征）的分析结果；3.3章节：特征权重分析（或有）及最终入模特征清单。*

# 模型设计与开发



## 模型设计

*【说明】介绍建模过程中一共训练了哪些模型，分别使用了哪些算法，如何进行模型融合（或有），****相关算法原理写入附件一****。*

## 模型调优

*【说明】主要介绍实际建模过程中的参数设计，使用了哪些方法进行模型参数调优，提供调参流程及对应参数调优后对模型效果的提升率。*

## 模型变量校验（或有，主要针对逻辑回归）

### （一）合理性校验

*【说明】对每个变量的符号从经验和业务合理性上进行检查。若变量的符号，即其预测的趋势，与业务经验相矛盾，则不论其显著性水平如何都需要将该变量删除或替换。*

### （二）相关性和多重共线性

*【说明】由于变量的强相关性和多重共线会影响逻辑回归模型的准确度和稳定性。开发中通过检查模型变量的相关系数和方差膨胀因子（VIF）等指标，以消除严重的多重共线性，一般标准为：*

* *相关系数的绝对值应不超过70%（对角线除外）；*
* *方差膨胀因子（VIF）均应<=5;*
* *模型变量矩阵的特征值显著不为0；*
* *条件指数均应<=10;*
* *变量的偏差比例在同一行出现较大值，表明这些变量具有较强相关性。*

*模型入模变量的相关系数和VIF检验结果如下表：*

### （三）特征分布

*【说明】入模变量应具有比较好的单调性，且其总体的好坏分布趋势与业务经验和模型一致。*

# 模型验证

*【说明】模型表现的评价指标及业界标准，详细信息可参考附件四。*



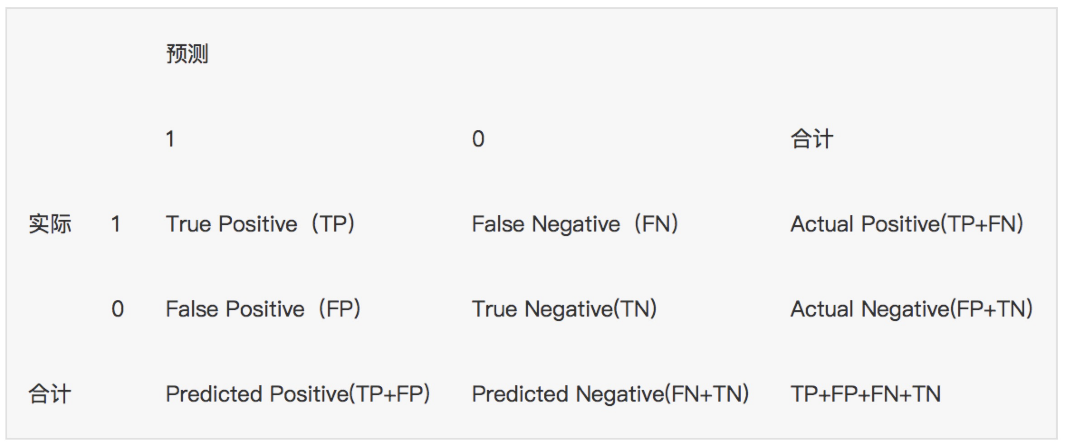
## 模型验证方法

*【说明】介绍模型验证策略，包括使用的数据集、验证方法、评估指标等。*

## 模型区分/排序能力指标

*【说明】提供模型在测试集上的混淆矩阵，召回率、精确率、F1、KS、ROC曲线、AUC、信息值、分离度等。另外根据实验业务需求提供其他模型衡量指标，如TOP N的模型召回率。*

1. *混淆矩阵、召回率、精确率和F1*

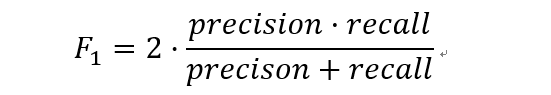


*如上图所示：TP（实际为正预测为正），FP（实际为负但预测为正），FN（实际为正但预测为负），TN（实际为负预测为负）*

*召回率（recall，查全率）： TP / (TP+FN)，预测对的正例数占真正的正例数的比率。*

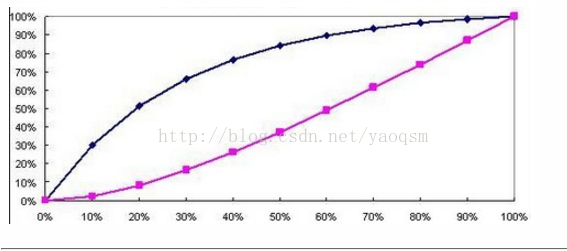
*精确率（precision，查准率）： TP / (TP+FP)，预测正确的正例数占预测为正例总量的比率。*

*F1：查准率和查全率的加权调和平均数*



1. *KS*

*如下图所示：KS曲线的纵轴是表示TPR和FPR的值，就是这两个值分别通过两条曲线来体现，横轴就是阈值（把0-1之间等分N份，等分点为阈值），然后在两条曲线分隔最开的地方，对应的就是最好的阈值，也是该模型最大的KS值。KS值的取值范围是[0，1] ，KS值越大，表示模型能够将正、负客户区分开的程度越大。*



*TPR=正例分对的概率 = TP/(TP+FN)，其实就是查全率*

*FPR=负例分错的概率 = FP/(FP+TN)，俗称假警报率*

1. *ROC曲线和AUC*

*ROC曲线及AUC系数主要用来检验模型对样本进行正确排序的能力。ROC曲线描述了在一定累计正样本比例下的累计负样本的比例，模型的分别能力越强，ROC曲线越往左上角靠近。AUC系数表示ROC曲线下方的面积。AUC系数越高，模型的区分能力越强，在下图中，AUC系数表示ROC曲线下方的面积。*



1. *信息值(Information Value, IV) 信息值是一个以熵为基础的区分能力指*

*标。信息值评估各等级或得分中违约户和非违约户分布的差别。信息值更多应用于零售评分卡变量设计过程中，在进行变量细分组到粗分组的过程中，评分卡变量的划分应以信息值较大的粗分组为最终划分结果。*

*（五）分离度（Divergence）是用于验证评分卡模型区分能力分离度。*

## 模型准确性/审慎性指标（一般用于评级模型、评分分段）

*【说明】*

1. *二项检验：验证内部评级体系各级别违约概率估计值准确性的一种有效方*

*法，其前提是每个级别下违约事件是相互独立的，在应用时应注意每次只能应用于一个评级级别。二项检验基于的假设是评级等级中的各个违约事件之间是独立的。*

1. *正态（Z）分布检验/t检验分布由于统计上的大数定理及中心极限定理，数*

*据足够多且方差有限的情况下，无论该数据的实际分布如何，该数据的均值的分布趋近于正态分布。因此，正态检验是应用非常广泛的统计检验手段。*

1. *条件信息熵(Conditional Information Entropy Ratio, CIER) 信息熵检*

*验主要是验证模型对违约对象的区分能力，通常是在模型预测全体样本的违约概率后，将全体样本按照评分划分为多个区间，然后用条件信息熵比率检验客户样本违约概率的分布是否有明显的异质性。*

1. *Brier得分（Brier Score， BS）及均方误差（Mean Squared Error, MSE）。*

*Brier得分与均方误差基于相同的理论。Brier得分检验评级模型预测的PD与二值变量违约与否（违约为1，非违约为0）的均方差。对于LGD和CCF模型，均方误差，MSE，是衡量估计量与被估计量之间差异程度的指标。*

## 模型稳定性指标

*【说明】通过比较开发时点与跨时间验证时点的样本预测结果，检查结果在跨时间时点上是否有较大波动。例如在评分卡的监控和验证中，通过群体稳定性指标（PSI），来确定人群的稳定性。针对模型稳定性分析结果也可以单独形成分析报告作为附件材料。评级模型考虑迁移矩阵，评分卡CSI。*

## 评分分布矩阵（或有）

*【说明】评分分段与好坏客群的二维交叉分布矩阵，或模型评分与已有评分的二位交叉矩阵，用以查看评分对于整个客户分布的区分能力。可以附加excel。*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分 | 好客户数 | 好客户占比 | 累计好客户数 | 好客户累计占比 | 坏客户数 | 坏客户占比 | 累计坏客户数 | 坏客户累计占比 | 好坏比 | 客户数 | 总客户占比 | 累计客户数 | 累计占比 |
| 660分以上 | 1000 | 10% | 1000 | 10% | 20 | 1% | 20 | 1% | 50 | 1020 | 5% | 1020 | 5% |
| 600-660（含） | 2000 | 20% | 3000 | 30% | 40 | 2% | 60 | 3% | 50 | 2040 | 10% | 3060 | 15% |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |

## 集中度

*【说明】赫芬达尔指数，简称“HHI”，是用来衡量资产组合集中度的常用指标之一。定义为各组合内客户数量比重的平方之和，或者各敞口比重的平方之和。赫芬达尔指数可用来验证模型评级结果精细程度，使用时应注意确保验证样本在验证期间具备对总体样本的代表性。*

*评分模型，某一分段客户占比不能太过集中，没有具体业界量化标准。*

# 模型优化与迭代（或有）

*【说明】说明模型优化和迭代的策略和方案，主要包括迭代训练模型的周期、否考虑引入新的特征变量，衍生更丰富的特征、是否应用新的算法优化模型等等。*

*也可以尝试分阶段进行，参考优化迭代的流程图如下：*



模型优化迭代流程图

# 附件一 算法概述



# 附件二 数据质量分析报告



# 附件三 特征工程设计



# 附件四 模型评价指标



# 附件五 模型评价指标业界标准

