天善智能
TIANSHAN SOFT

申请评分卡模型

数据的预处理与特征构建

讲师：安迪生

课程简介

- 在构建评分卡模型的工作中，数据预处理工作和特征构建工作是至关重要的一步。数据的预处理工作可以有效处理缺失值与异常值，从而增强模型的稳健性。而特征构建工作则可以讲信息从字段中加以提炼，形成有业务含义的优异特征。

目录

评分卡模型的简介

- 什么是评分卡

风控场景中的评分卡

✓以分数的形式来衡量风险几率的一种手段

✓是对未来一段时间内违约/逾期/失联概率的预测

✓有一个明确的(正)区间

✓通常分数越高越安全

✓数据驱动

✓反欺诈评分卡、申请评分卡、行为评分卡、催收评分卡

非信贷场景中的评分卡

✓推荐评分卡

✓流失评分卡

评分卡模型的简介

- 常用的信贷评分卡

申请评分卡（Application Scorecard）

用在贷前审核环节，评估放贷后是否会违约的模型。常用特征：个人信息、央行征信信息、申请行为信息、其他辅助信息

行为评分卡（Behavioral Scorecard）

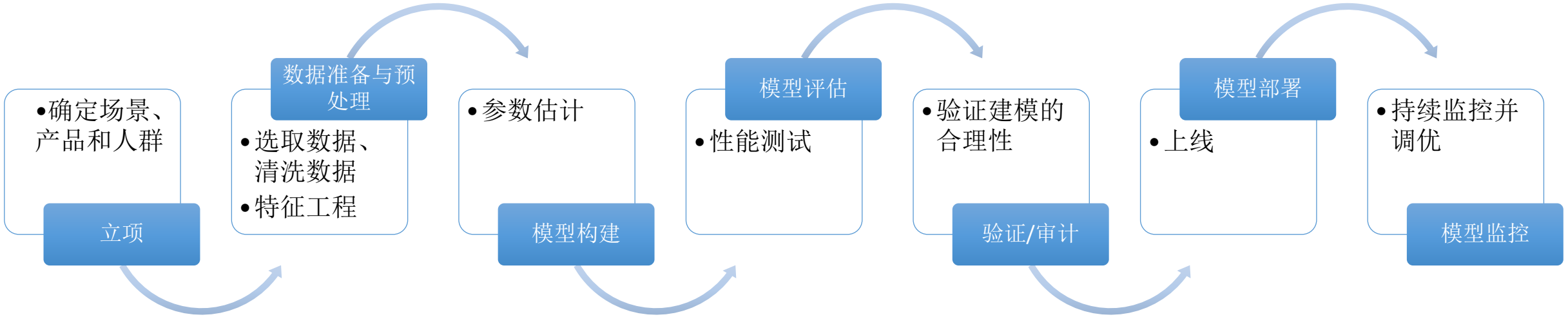
用在贷后监控环节，做早期预警的工作（包括巴塞尔2.5及之前的AIRB的要求）。常用特征：贷后的还款行为、消费行为等。通常适用于还款周期长的产品或者循环授信类产品

催收评分卡（Collection Scorecard）

用在发生逾期后的管理环节，为催收工作提供指导。催收评分卡又可细分为预测失联的失联评分卡、预测逾期加重的滚动率评分卡和预测催收后的还款率的还款率评分卡。常用特征：个人信息、贷后的还款行为、消费行为、联系人信息等

评分卡模型的简介

- 评分卡模型开发步骤



评分卡模型的简介

- 评分卡开发的常用模型

逻辑回归

优点: 简单，稳定，可解释，技术成熟，易于监测和部署

缺点：准确度不高

决策树

优点: 对数据质量要求低，易解释

缺点：准确度不高

其他元模型

组合模型

优点: 准确度高，不易过拟合

缺点：不易解释；部署困难；计算量大

数据集介绍

- 数据集介绍

本次案例分析用的数据，是拍拍贷发起的一次与信贷申请审核工作相关的竞赛数据集。其中共有3份文件：

- PPD_Training_Master_GBK_3_1_Training_Set.csv：信贷客户在拍拍贷上的申报信息和部分三方数据信息，以及需要预测的目标变量
- PPD_LogInfo_3_1_Training_Set.csv：信贷客户的登录信息
- PPD_Userupdate_Info_3_1_Training_Set.csv：部分客户的信息修改行为

建模工作就是从上述三个文件中对数据进行加工，提取特征并建立合适的模型，对贷后表现做预测。

数据集介绍

- 数据集介绍（续）

表1：PPD_Training_Master_GBK_3_1_Training_Set

Idx	UserInfo_1	UserInfo_2	UserInfo_3	UserInfo_4	WeblogInfo	WeblogInfo	WeblogInfo	ThirdParty	ThirdParty	ThirdParty	ThirdParty	ThirdParty	ThirdParty	target	ListingInfo
10001	1	深圳	4	深圳	0	1	0	10	47	167	0	25	65	0	2014/3/5
10002	1	温州	4	温州	0	1	0	0	0	68	105	40	2	0	2014/2/26
10003	1	宜昌	3	宜昌	0	1	0	1	2	50	50	49	0	0	2014/2/28
10006	4	南平	1	南平	0	1	0	9	54	56	125	38	0	0	2014/2/25
10007	5	辽阳	1	辽阳	0	0	0	1	5	39	34	36	73	0	2014/2/27
10008	1	吴忠	5	银川	0	0	0	31	43	233	205	80	3	0	2014/2/27
10011	1	绵阳	3	赤峰	0	1	0	6	3	64	29	33	93	1	2014/2/24
10015	4	东莞	5	东莞	0	0	0	-1	-1	-1	-1	-1	-1	0	2014/2/28
10019	1	赤峰	6	赤峰	0	1	0	180	3	414	363	85	0	1	2014/2/24
10021	3	武汉	5	鄂州	0	0	0	14	28	132	118	73	2	0	2014/2/27
10022	5	武汉	5	武汉	0	1	0	22	0	163	165	83	3	0	2014/2/25
10024	5	长沙	5	长沙	0	1	0	14	9	90	149	51	2	0	2014/3/6
10026	3	漳州	3	漳州	1	1	0	0	8	2	22	6	0	0	2014/3/4
10027	3	牡丹江	6	牡丹江	0	1	0	49	37	363	130	123	5	1	2014/2/26
10031	1	太原	3	太原	0	1	0	87	35	352	467	106	7	0	2014/3/7
10032	1	北京	6	北京	0	1	0	12	14	162	242	146	4	0	2014/2/26

表2：PPD_LogInfo_3_1_Training_Set

Idx	Listinginf	LogInfo1	LogInfo2	LogInfo3
10001	2014/3/5	107	6	2014/2/20
10001	2014/3/5	107	6	2014/2/23
10001	2014/3/5	107	6	2014/2/24
10001	2014/3/5	107	6	2014/2/25
10001	2014/3/5	107	6	2014/2/27
10001	2014/3/5	107	6	2014/3/4
10001	2014/3/5	1	1	2014/2/20
10001	2014/3/5	1	20	2014/2/20
10001	2014/3/5	12	0	2014/2/20
10001	2014/3/5	1	2	2014/2/20
10001	2014/3/5	2	1	2014/2/20
10001	2014/3/5	4	1	2014/2/20
10001	2014/3/5	-4	6	2014/2/20
10001	2014/3/5	-4	6	2014/2/20
10001	2014/3/5	-4	6	2014/2/23
10001	2014/3/5	-4	6	2014/2/24

表3：PPD_Userupdate_Info_3_1_Training_Set

Idx	ListingInfo1	UserupdateInfo1	UserupdateInfo2
10001	2014/3/5	_EducationId	2014/2/20
10001	2014/3/5	_HasBuyCar	2014/2/20
10001	2014/3/5	_LastUpdateDate	2014/2/20
10001	2014/3/5	_MarriageStatusId	2014/2/20
10001	2014/3/5	_MobilePhone	2014/2/20
10001	2014/3/5	_MobilePhone	2014/2/20
10001	2014/3/5	_QQ	2014/2/20
10001	2014/3/5	_ResidenceAddress	2014/2/20
10001	2014/3/5	_ResidencePhone	2014/2/20
10001	2014/3/5	_ResidenceTypeId	2014/2/20
10001	2014/3/5	_ResidenceYears	2014/2/20
10002	2014/2/26	_age	2013/6/21
10002	2014/2/26	_educationId	2013/6/21

数据集介绍

- 关键字段

数据源	字段	含义
PPD_Training_Master_GBK_3_1_Training_Set	Idx	用户的唯一标识
	Target	目标变量，以1、0表示违约与非违约
	UserInfo_*	用户属性，多以归属地内容，有缺失
	ThirdParty_Info_Period*	除-1外，其他都是非负整数。-1可能是特殊值
	ListingInfo	贷款发放日期
PPD_LogInfo_3_1_Training_Set	LogInfo3	登录日期
	LogInfo*	登录事件代码
PPD_Userupdate_Info_3_1_Training_Set	UserupdateInfo1	更改信息的所属字段
	UserupdateInfo2	更改日期

特征构造

- 特征构造的重要性

在评分卡模型的开发中，特征构造是极其关键的步骤，其作用是将分散在不同字段中的信息加以组合，从中提炼出有价值的、可用的信息进而进行评分卡模型的开发。

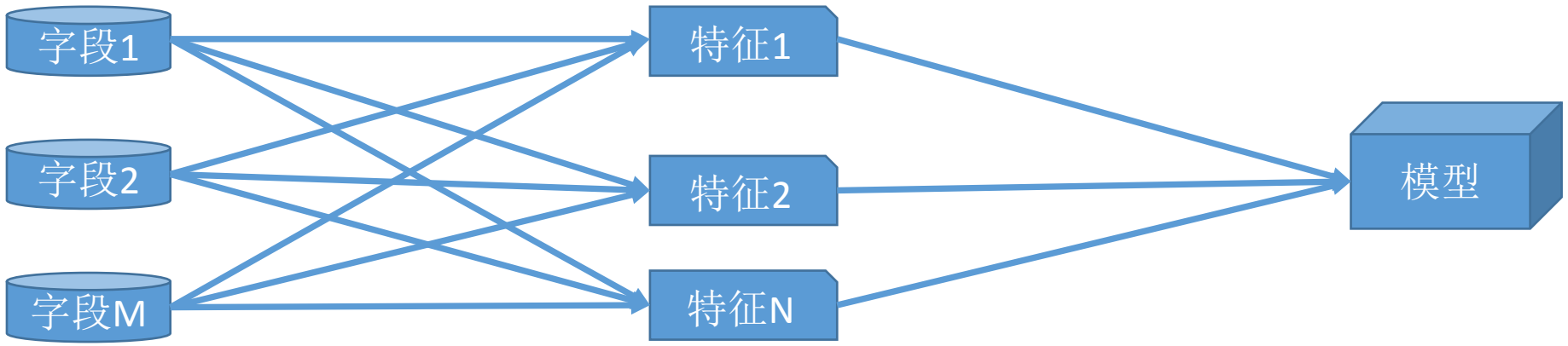
部分常用的特征构方法有：

求和：例如过去一段时间内的每月网购金额的总和

比例：例如申请贷款的月还款本息与月收入的占比

频率：例如过去一段时间内的境外消费次数

平均：例如过去一段时间内平均每次信用卡取现额度



特征构造

- 好的特征需要具备以下优势

稳定性高

当人群分布稳定、产品营销稳定、宏观经济因素稳定、监管政策稳定时，特征的分布也需要稳定

区分度高

未来的违约与非违约人群在特征上的分布需要显著不同

差异性大

不能对全部人群或绝大部分人群上有单一的取值

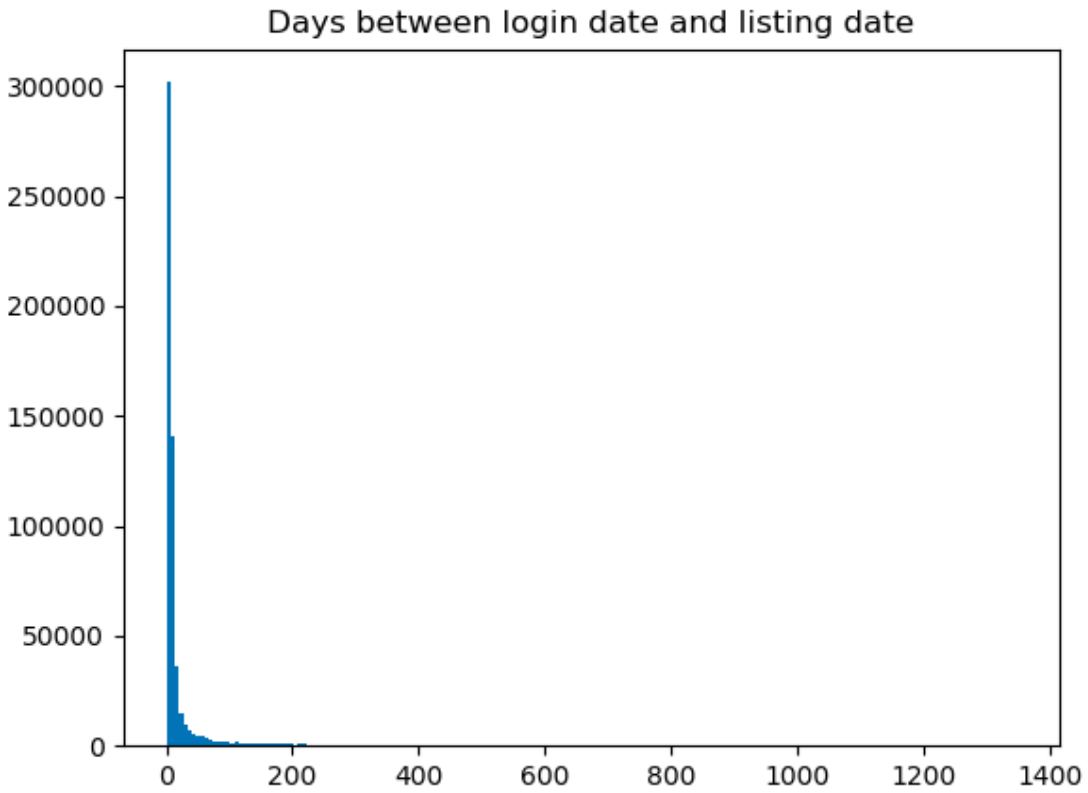
符合业务逻辑

特征与信用风险的关联关系要符合风控业务逻辑

特征构造

- 案例：对PPD_LogInfo_3_1_Training_Set字段的处理

在该数据源中，我们有代表身份的idx、代表登录日期的LogInfo3和操作代码LogInfo1与LogInfo2。计算登录日期与放款日期之间的间隔天数，可以看到绝大部分的天数在180天以内.



注意到，虽然LogInfo1与LogInfo2取值为数值，但是其含义是类别。因此这类变量不能求和、平均值、最值等。但是可以求频率和个数。

此外，我们需要考虑到间隔天数在变量构造中的作用。例如，对于计算某种操作的频率，我们可以考虑近30天内、近60天内、近90天内等等不同的时间窗口，称之为时间切片。

时间切片的设定：
不宜太长，否则大部分样本的时间跨度无法满足
不宜太短，否则抓取不到足够多的信息，且变量不稳定

特征构造

- 案例：对PPD_LogInfo_3_1_Training_Set字段的处理（续）

由于绝大部分观测样本的时间跨度在半年内，所以我们选取半年内的时间切片，考虑以月为单位的时间切片，则可以衍生出30天、60天、90天、120天、150天、180天等多种选择。

同时，对于类别型变量，可以考虑构造如下计算逻辑：

时间切片内的登录的次数

时间切片内不同的登录方式的个数

时间切片内不同登录方式的平均个数

不同的时间切片与不同的计算逻辑的交互可以产生多个特征。这些特征往往存在一定程度上的线性相关性。在接下来的多变量分析中，需要消除线性相关性对模型产生的影响。

注意：

该数据源中，每个idx存在多条记录。上述的特征构造是针对每个idx进行相应的计算。

数据的质量检验

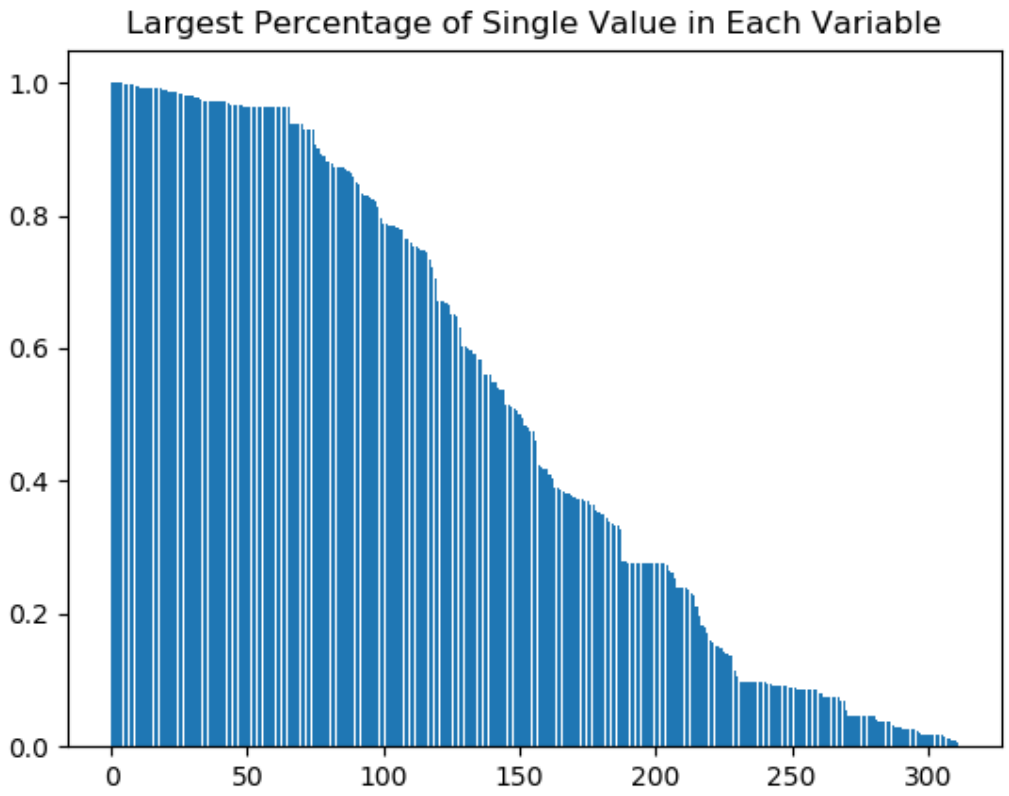
• 数据的质量检验-数据集中度

在信用风控模型的开发中，数据集中度是常见的问题。即在变量中，某单一数值的占比就占了全部样本值的绝大多数。例如，在一批训练样本中，学历为本科的样本占了全部样本的90%。具有极高的集中度的字段或者变量，需要按照风险程度进行区分：

1. “多数值”与“少数值”对应的坏样本率没有显著差别。此时，由于包含的信息较少、对模型的开发没有太大价值，而且往往“少数值”的产生是由于误差或者噪声，我们可以直接将字段删除。

2. “多数值”与“少数值”有显著差别，且“少数值的坏样本率低于“多数值”。此时，尽管二者的差别很显著，但是由于我们更加关注风险度高的一组所以“少数值”的存在并不会带来额外的意义，此时也可以直接将字段删除。

3. “多数值”与“少数值”有显著差别，且“少数值的坏样本率高于“多数值”。此时，“少数值”的存在表明该值对应的风险很高，字段需要保留。



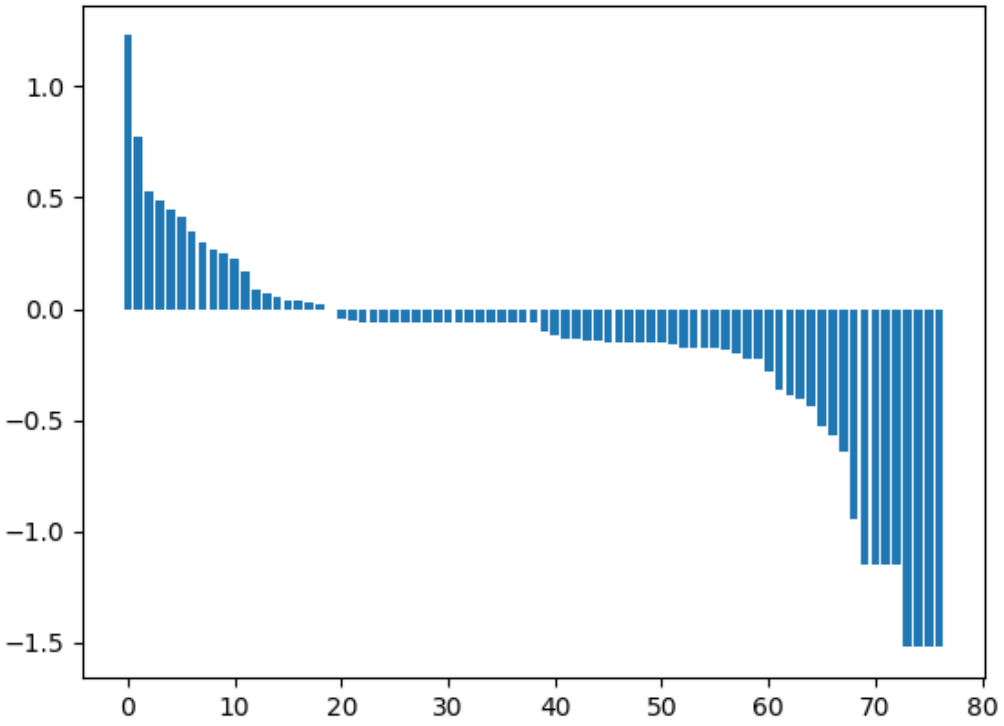
天善智能
TIANSHAN SOFT

数据的质量检验

• 数据的质量检验-数据集中度（续）

本案例在衡量两组样本的坏样本率的差异时，用二者的坏样本率的比值的对数作为衡量。对数转换可以将互为倒数的两组数变为互为相反数。例如，当A为B的10倍时，取对数后为2.303。当A为B的十分之一时，取对数后为-2.303。我们重点考察上一页中的第3点，即“少数值”的坏样本率是否高于“多数值”，以对数比例2.303为界限。

我们检查多数值占比超过90%的字段，对每个字段都检查多数值的坏样本率与少数值的对数比例。计算后发现，所有的对数比例都在2.303以内，说明少数值的存在并不会对风险产生影响，因此可以将这部分字段删除。



数据的质量检验

• 数据的质量检验-数据缺失（data missing）

数据缺失度是数据质量检验的一个重要项。需要从两个维度检验数据缺失度：

✓字段维度，即某个字段在全部样本上的缺失值个数的占比

✓样本维度，即某条样本在所有字段上的缺失值的占比

一般而言，字段维度的缺失程度会大于样本维度的缺失程度

• 缺失值处理

➢舍弃该字段或该条记录：缺失占比太高

➢补缺：缺失占比不高，可用均值法、众数法、回归法等

➢作为特殊值：将缺失看成一种特殊值

数据的质量检验

- 数据的质量检验-数据缺失（续）

其中，补缺的方法依变量类型的不同而有所差异。比如，均值法和回归法适用于数值型变量，众数法适用于类别型变量。我们需要分辨出变量是属于类别型还是数值型。在实际业务中可按照下述的准则来判断变量的类型：

- 当且仅当变量取值为数值，且不同值的个数比较多时，视为数值型变量，这时可以用均值法（完全随机缺失）、抽样法（完全随机缺失）、回归法（针对随机缺失）进行补缺
- 其他情况下均视为类别型变量，这时可以用抽样法、众数法进行补缺。

补缺工作的前提是，字段整体的缺失率不宜太高，否则会产生较大的偏差且对字段的使用（包括由该字段衍生的特征）的使用效果产生影响。

数据的质量检验

• 数据的质量检验-数据缺失（续）

在信贷评分模型中，数据的缺失包含着多重意义。很多时候是完全非随机缺失，其缺失状态有着业务含义。例如，某些信贷产品的申请环节需要提供芝麻分，而且该字段的缺失本身对应的风险就比较高。有些时候缺失是完全随机缺失，缺失与否并不影响信用风险。对于不同的缺失机制，对应的处理方法也有所不同。

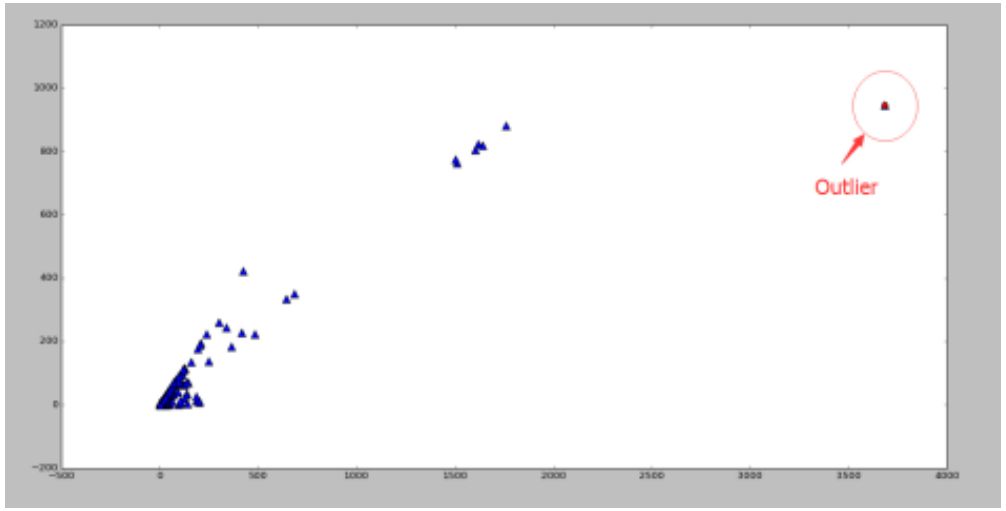
- 完全非随机缺失：有缺失值的样本的违约率显著高于无缺失样本，此时应当将缺失当成一种特殊的状态

- 完全随机缺失：有缺失值的样本的违约率与无缺失样本无明显差异，此时如果缺失样本的占比很少，可将样本删除。如果缺失样本的占比较高，需要将字段删除。

数据的质量检验

- 数据的质量检验-异常值（outliers）

与缺失值类似，异常值在一般的数据分析场景中也会对模型产生一定的干扰，需要对其做处理。异常值的判断通常有聚类法、分位点法等等，处理方法有删除法、替换法。



$$x > Q_3 + 3(Q_3 - Q1)\ or\ x < Q_1 - 3(Q_3 - Q_1)$$

Q_3, Q_1分别是样本的75%与25%分位点

数据的质量检验

- 数据的质量检验-异常值（outliers）

但是在信用评分模型中，异常值往往也带有特殊的意义，例如，在提交的申请资料中，如果PBOC征信记录查询次数过多，可能该申请人在一定时间内申请贷款的次数过多，则很有可能该申请人面临的资金需求很迫切，对未来的逾期概率产生不好的影响。对于这部分人，在数据预处理阶段是不宜直接删除或者用正常值进行替换。评分卡模型的开发中，也有相应的方法来处理这样的异常值。

数据的质量检验

- 数据的质量检验-数据含义一致性

在实际工作中，数据的录入中往往会使得原本属于同一含义的记录值出现不同的记录。例如，通讯方式“QQ”与“qq”是一类性质，或者手机号码“+8613000000000”与“13000000000”均表示同一个号码。因此，我们需要将具有相同含义的数据进行统一。

本案例中，需要手动地将“QQ” 和“ qQ”，“Idnumber” 和“ idNumber”以及“MOBILEPHONE”和“PHONE”进行统一。

更多商业智能BI和大数据精品视频尽在 www.hellobi.com

BI、商业智能
数据挖掘 大数据
数据分析师
R语言 Python
机器学习
深度学习
人工智能
Hive Hadoop
Tableau
BIEE ETL
数据科学家
PowerBI

天善智能
TIANSHAN SOFT