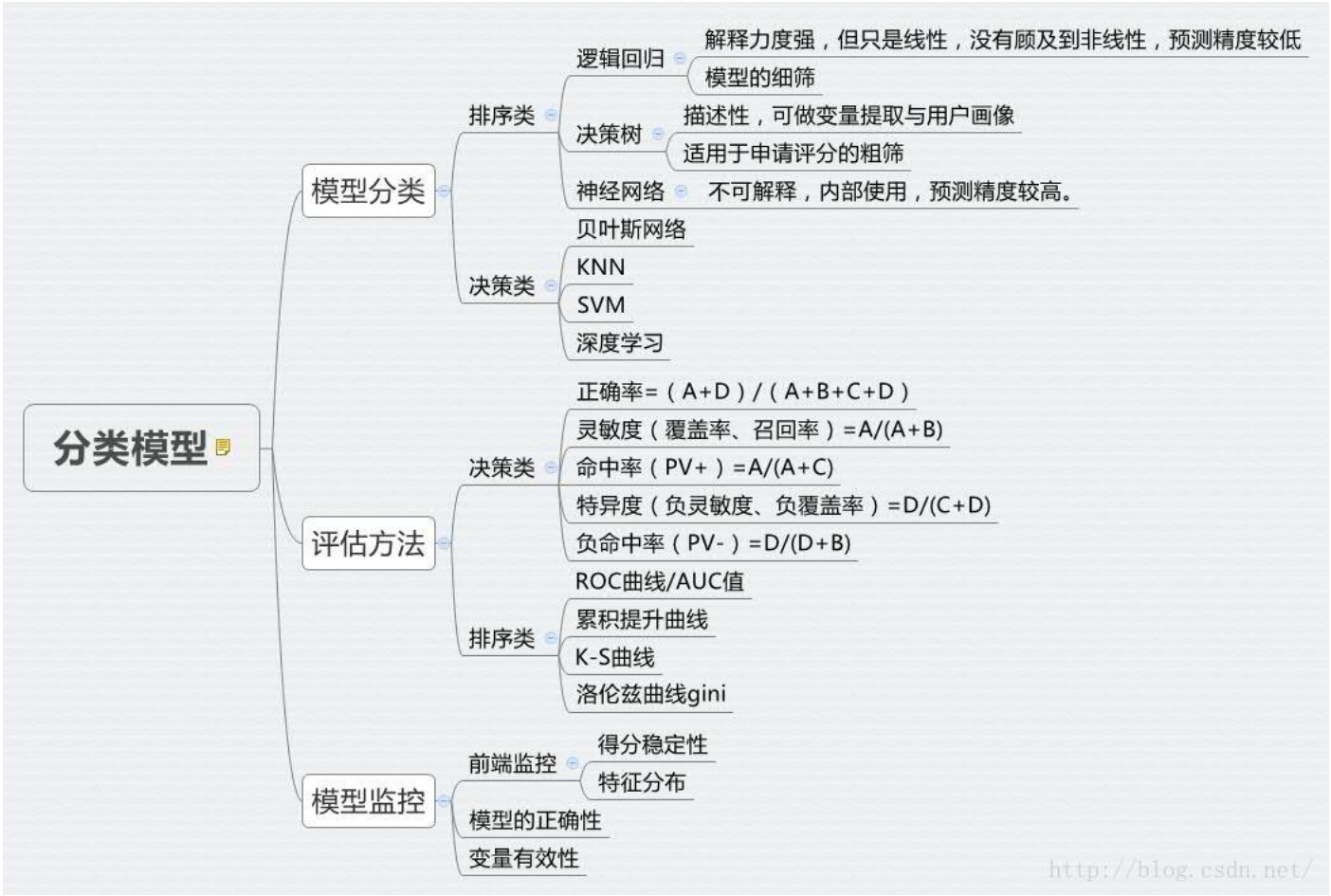
笔记丨风控分类模型种类（决策、排序）比较与模型评估体系（ROC/gini/KS/lift）

原创　2016年06月21日 10:56:42　19194

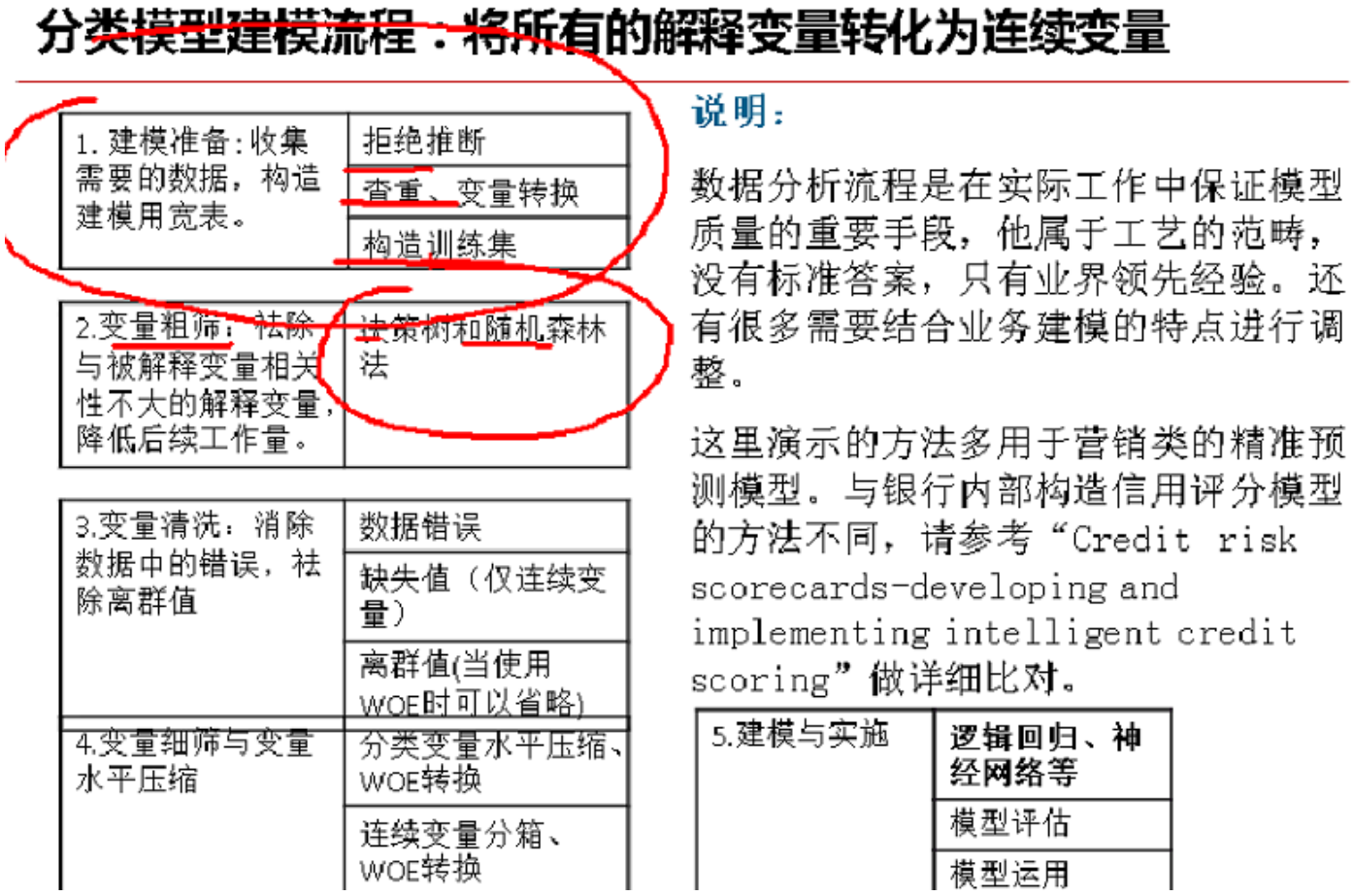
本笔记源于CDA-DSC课程，由常国珍老师主讲。该训练营第一期为风控主题，培训内容十分紧凑，非常好，推荐：CDA数据科学家训练营 (http://www.cda.cn/dsc/)



一、风控建模流程以及分类模型建设

1、建模流程

该图源自课程讲义。主要将建模过程分为了五类。数据准备、变量粗筛、变量清洗、变量细筛、建模与实施。



2、分类模型种类与区别

风控与其他领域一样，分类模型主要分为两大类：排序类、决策类、标注类（文本、自然语言处理）。

预测类型	方法	适用场景	举例
排序（Rankings）	逻辑回归 决策树 神经网络	不存在稳定的可辨识的结果。比如流失经常是一个定义，而很少存在真实流失的情况	信用评分 流失预测 营销响应
决策（Decisions）	贝叶斯网络、KNN（基于记忆的模型）、SVM、深度学习	存在可以直接辨识的结果。比如人脸图像识别，是可以直接知道是否为某个人的脸	声音识别 图像识别 欺诈识别
标注（Tagging）	隐马尔可夫 条件随机场	存在明确的分类，和决策的不同在于决策为二分类，标注为多分类	信息抽取 自然语言处理

一般来说风控领域在意的是前两个模型种类，排序类以及决策类。

其中：**巴塞尔协议定义了金融风险类型：市场风险、作业风险、信用风险。信用风险ABC模型有进件申请评分、行为评分、催收评分。**

模型	解释	复杂度
Logistics回归	影响程度大小与显著性，解释力度强，但只是线性，没有顾及到非线性，预测精度较低	
决策树	1、描述性，重建用户场景，可做变量提取与用户画像	叶子的数量

	2、树的结构不稳定，可以得出变量重要性，可以作为变量筛选	
随机森林	随机森林比决策树在变量筛选中，变量排序比较优秀	
神经网络	1、不可解释，内部使用，预测精度较高。可以作为初始模型的金模型（用以评估在给定数据条件下，逻辑回归可达到的最精确程度） 2、线性（逻辑回归）+非线性关系，可用于行为评分的预测模型（行为评分对模型可解释性不强），可用于申请评分的金模型 3、使用场景：先做一个神经网络，让预测精度（AUC）达到最大时，再用逻辑回归	迭代次数

(1)信用风险——申请信用评分

申请评分可以将神经网络+逻辑回归联合使用。
《公平信用报告法》制约，强调评分卡的可解释性。所以初始评分（申请评分）一般用回归，回归是解释力度最大的。
神经网络可用于银行行为评级以及不受该法制约监管的业务（P2P）。其次，神经也可以作为申请信用评分的金模型。

金模型的使用：一般会先做一个神经网络，让预测精度（AUC）达到最大时，再用逻辑回归。
建模大致流程：
一批训练集+测试集+一批字段——神经网络建模看AUC——如果额定的AUC在85%，没超过则返回重新筛选训练、测试集以及字段；
超过则，可以后续做逻辑回归。

（2）信用风险——行为评分

行为评分建模：行为信用评级不需要解释性，所以可以用非线性的神经网络。

二、分类模型评估体系

上述将分类模型做了归纳，不同的分类模型所采用的评估体系不同。

决策类：准确率/误分率、利润/成本
排序类：ROC指标（一致性）、Gini指数、KS统计量、提升度

1、决策类评估——混淆矩阵指标

混淆矩阵，如图：其中这些指标名称在不同行业有不同的名称解释

混淆矩阵： 给定一个阈值，就可以做出一个混淆矩阵		打分值		
		反应（预测=1）	未反应（预测=0）	合计
真实结果	呈现信号（真实=1）	A（击中） True Positive	B（漏报） False Negative	A+B
	未呈现信号（真实=0）	C（虚报） False Positive	D（正确否定） True Negative	C+D
合计		A+C	B+D	A+B+C+D

正确率=（A+D）/（A+B+C+D）
灵敏度（覆盖率、召回率）=A/(A+B)
命中率（PV+）=A/(A+C)

特异度（负灵敏度、负覆盖率）=D/(C+D)
负命中率（PV-）=D/(D+B)

在以上几个指标中不同行业看中不同的指标：

（1）灵敏度/召回率/覆盖率（——相对于命中率）

譬如灵敏度（召回率）这一指标就比正确率要重要，覆盖率(Recall）这个词比较直观，在数据挖掘领域常用。因为感兴趣的是正例(positive)，比如在信用卡欺诈建模中，我们感兴趣的是有高欺诈倾向的客户，那么我们最高兴看到的就是，用模型正确预测出来的欺诈客户(True Positive)cover到了大多数的实际上的欺诈客户，覆盖率，自然就是一个非常重要的指标。

（2）命中率（——相对于覆盖率）

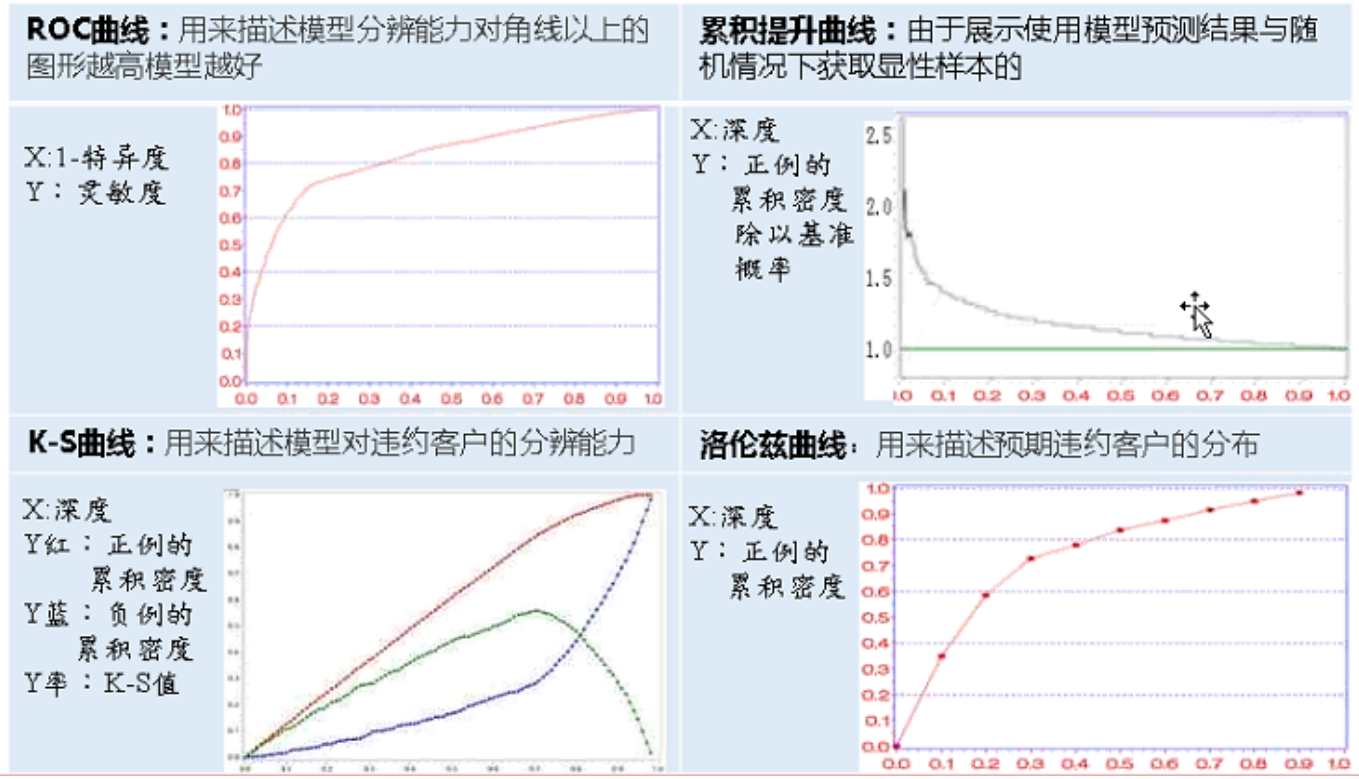
欺诈分析中，命中率（不低于20%），看模型预测识别的能力。
在数据库营销里，你预测到b+d个客户是正例，就给他们邮寄传单发邮件，但只有其中d个会给你反馈（这d个客户才是真正会响应的正例），这样，命中率就是一个非常有价值的指标。 以后提到这个概念，就表示为PV+(命中率，Positive Predicted Value)*。

2、排序类指标评估

ROC指标（一致性）、Gini指数（洛伦兹曲线）、KS统计量、提升度四类指标。

排序类模型的评估指标

该类模型的需求是回答“会不会？”。比如预测一下客户违约的概率、营销响应的概率



（1）ROC曲线

对角线模型，最差，风控喜欢的指标。由决策类指标的灵敏度（召回率/覆盖率）与特异度（负灵敏度、负召回率）来构造。

求覆盖率等指标，需要指定一个阈值（threshold）。随着阈值的减小，灵敏度和1-特异度也相应增加（也即特异度相应减少）。

把**基于不同的阈值而产生的一系列灵敏度和特异度描绘到直角坐标上，就能更清楚地看到它们的对应关系。**把sensitivity和1-Specificity描绘到同一个图中，它们的对应关系，就是传说中的ROC曲线，全称是receiver operating characteristic curve，中文叫“接受者操作特性曲线”。

AUC值，为了更好的衡量ROC所表达结果的好坏，Area Under Curve（AUC）被提了出来，简单来说就是曲线右下角部分占正方形格子的面积比例。该比例代表着分类器预测精度。（R语言丨ROC曲线——分类器的性能表现评价 (http://blog.csdn.net/sinat_26917383/article/details/51114244)）

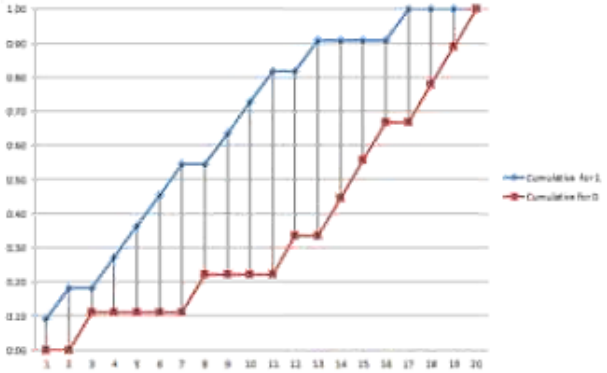
（2）累积提升曲线

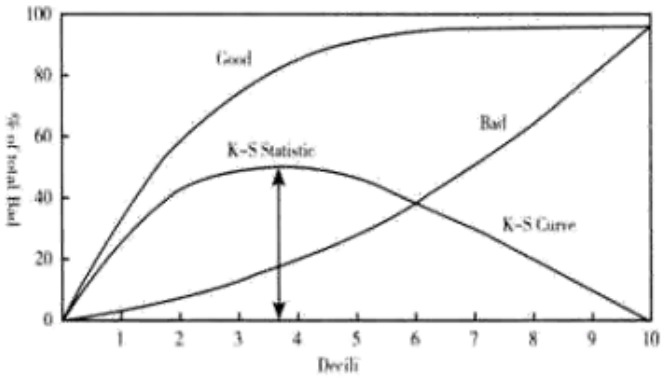
营销最好的图，很简单。它衡量的是，与不利用模型相比，模型的预测能力“变好”了多少（分类模型评估——混淆矩阵、ROC、Lift等 (http://iflypig.com/?p=270)）。
将概率从大到小铺开x，提升度可以有一些“忽悠”的成本，哈哈~可以微调，可以自己调节提升度的区间

（3）K-S曲线

风控喜欢的指标。K-S曲线的最大值代表K-S统计量。

K-S曲线





K-S统计量

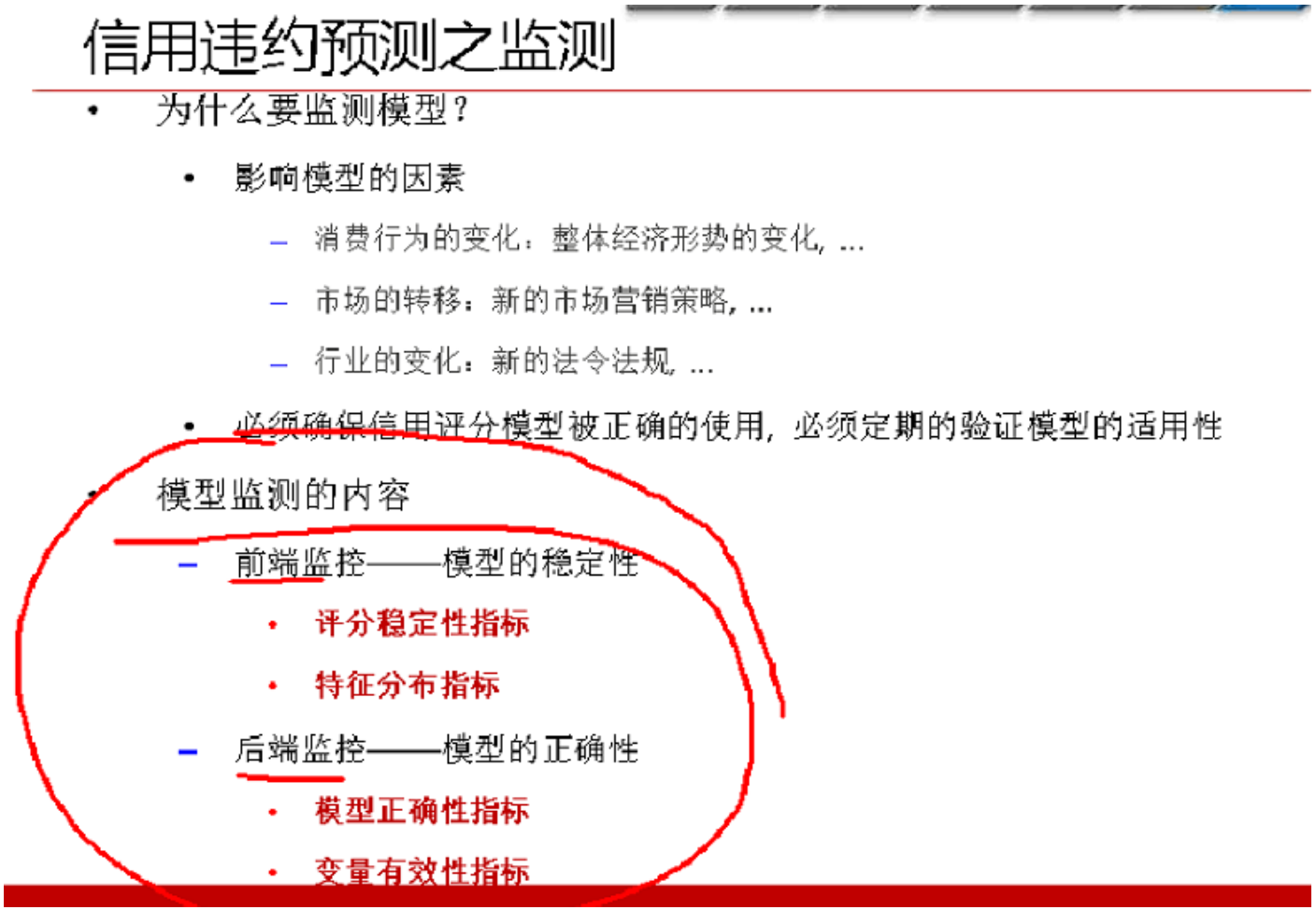
- ✓小于20：模型无鉴别能力
- ✓20~40之间：模型勉强接受
- ✓41~50之间：模型具有区别能力
- ✓51~60之间：此模型有很好的区别能力
- ✓61~75之间：此模型有非常好的区别能力
- ✓大于75：此模型异常，可能有问题

（4）洛伦兹曲线gini

风控喜欢的指标，TP率给了一个累积比，跟提升度差不多。

三、信用风险模型检测

监测可以分为前端、后端监控。



前端监控，授信之前，别的客户来了，这个模型能不能用？
后端监控，建模授信之后，打了分数，看看一年之后，分数是否发生了改变。

1、前端监控

长期使用的模型，其中的变量一定不能波动性较大。比如，收入这个指标，虽然很重要，但是波动性很大，不适合用在长期建模过程中。
如果硬要把收入放到模型之中，要放入收入的百分位制（排名）。

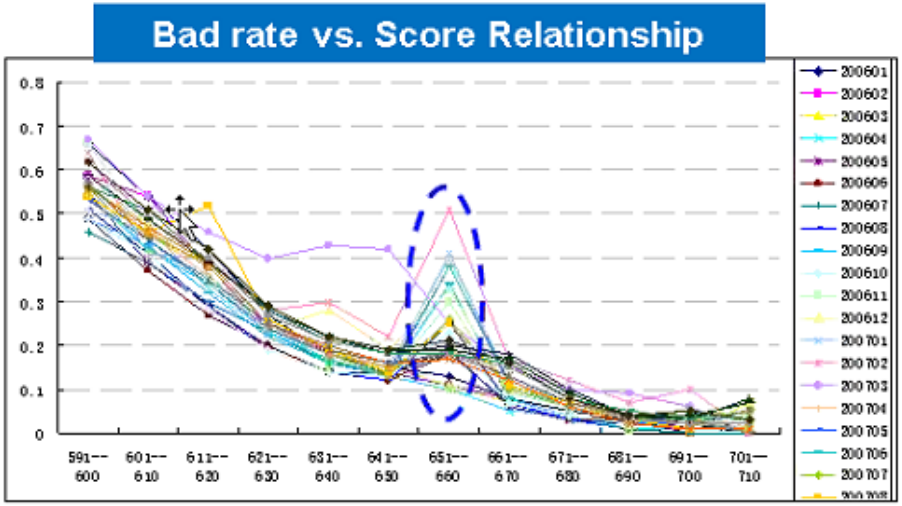
2、后端监控

主要监控模型的正确性以及变量选择的有效性。出现了不平滑的问题，需要重新考虑

信用违约预测之监测

后端监控——模型正确性

评分区间 651-660上行为评分模型对风险的区分能力较差



信用违约预测之监测

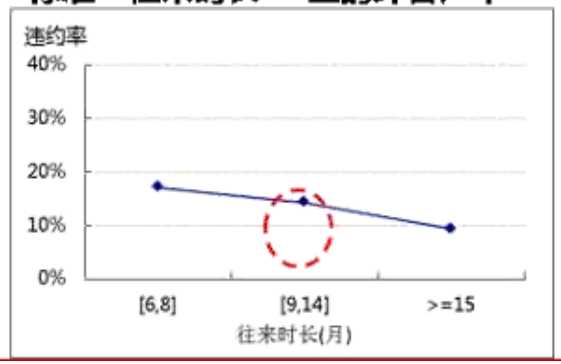
后端监控——变量有效性

行为评分卡

决策属性

客户属性	帐户属性	消费行为	还款行为
• 年龄	• 往来时长	• Recency	• 风险评价
• 居住状况	• 金卡/普卡	• Frequency	• 逾期频率
• 性别	• 信用额度	• Monetary	• 催收
• 婚姻状况	• 自动还款	• 额度使用率	• 循环信用
• ...	• ...	• ...	• ...

标准“往来时长” 上的坏客户率



实际“往来时长” 上的坏客户率