风控建模  
  
1. 什么是信用评分  
 信用评分就是衡量你的信用程度的量化指标，这样听起来是不是很拗口？说简单点，就是一个分数，  
 由你填写的个人信息以及一些第三方数据计算得来的，例如支付宝的芝麻信用、腾讯的腾讯信用、美国的FICO评分等。  
 这些评分就可以决定你的信用等级，从而让贷款机构决定放不放款  
  
2. 常用的模型  
 业界常说的有A卡、B卡、C卡，A卡就是申请评分卡。在你申请的时候就会站出来，决定放不放款，  
 B卡，也就是贷中行为评分卡，监控你的信用状况，决定给不给你提额度，或者中不中断你的贷款，  
 C卡就是贷后评分卡，一般有三种：账龄迁移模型、还款率模型和失联预警模型。  
   
 迁移模型、还款率模型和失联预警模型。   
 账龄迁移：就是预测你的逾期状况会不会从M1迁移到M2（关于m1，m2定义查看文末解释）   
 失联预警模型：对于银行和贷款公司来讲，有时不怕你不还钱，如果逾期了还能对你进行罚息等方式再赚一笔，更怕的反而客户失联，彻底消失，所以就需要建立失联预警，看一看你未来有没有可能失联。   
 还款率模型：注意这个模型不是为了预测你还不还钱，而是预测未来经过催收动作后，还款的概率。  
 毕竟贷后催收，是需要人力、物力成本的，如果一个人简单催一催，例如发个短信，打个电话就还了，  
 何必动用戴金链子的老铁们上门，出于平衡成本的需要，制定不同的催收套路，这个模型就可以起作用了。  
  
3 概念解析：M0,M1,M2的定义  
 这三个不是简简单单的月份的概念，具体定义如下：  
  
 M0:最后还款日的第二天到下一个账单日  
 M1:M0时段的延续,即在未还款的第二个账单日到第二次账单的最后还款日之间  
 M2:M1的延续,即在未还款的第三个账单日到第三次账单的最后还款日之间  
 所谓账单日，就是银行会对你的当期应还款形成账单并通知你，账单形成日即为账单日。，同时，银行不会要求你马上就还款，而是会给你一个缓冲期，通常是20天（具体根据各银行制定标准），该期限截止日即为还款到期日。这20天之内全额还款或是选择信用卡最低还款额方式还款，可以享受免息待遇，但如果逾期，就会计息了。  
  
 这些概念看起来是不是很难懂，这样，我举个例子你就明白了，就拿我的招行信用卡来说的吧。还款日是每个月4号，帐单日是每个月16号。   
 假如我在1月16号前消费了500块，那么一月帐单日，也就是1月16日账单金额就是500，还款到期日就是2月4号。缓冲期就是1月16日–2月4日，一共19天，如果在这期间我还清了账单，就不会有利息。   
 所谓的M0就是2月5号到下个月账单日—-3月16日期间我不还清账单，也就是产生了M0逾期。   
 假如我继续不还，也就是3月16号我还拖着，直接延伸到M1，那么M1就是3月16日到4月4日。   
 如果到了M1依然不还钱，第三个账单日就是4月16号，M2就是4月16日到5月4日。

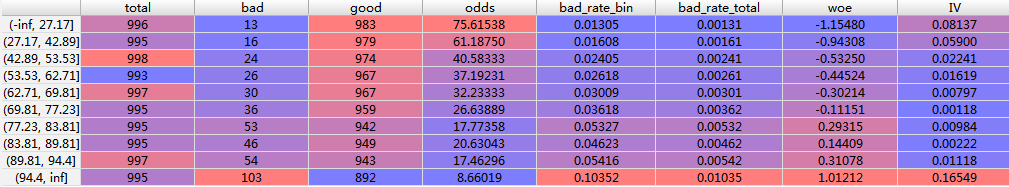
WOE：god(i)/god(sum)/bad(i)/bad(sum)  
具体参考小象课程第三课文档  
   
建模过程：  
 1. 数据读入  
 2. 数据概览  
 3. 数据转换  
 文本转数字  
 4. 数据处理  
 缺失值，异常值的处理  
 5. 数据切分  
 train，test，oot  
 6. 数据分箱  
 卡方分箱，数据转换WOE，VI值  
   
 7. 训练模型  
 选择模型，  
   
 8. 验证模型  
 AUC，KS ROC 等  
   
 9 信用评分  
   
分箱过程  
 在实际的工作中，特征数据情况复杂，按照数据的类型看，包含比如日期类型，文字类型，数值类型等变量，按照某一个变量的离散情况，又分为离散变量和连续变量。

bad rate坏样本率，指的是将特征进行分箱之后，每个bin下的样本所统计得到的坏样本率  
 我们一般定义分箱数为5箱，也就是说最后每一个变量的数据都转换为小于等于5类的数据分别放在5个分箱里面，并且每一个分箱里面的标签都同时包含了好坏样本。  
 1. 检查变量的数值类型，将大于5类小于5类的变量区分出来。  
 如果var<5时，分别遍历每一个var中的每一类数据是否都同时包含好换样本。  
 如果 某一个变量的某一类数据所对应的标签全是坏的或好的。  
 合并这个变量其中的两类数据，进而保证同时包含好坏样本，比如v1这个变量有3种数值1，2,3 但是1这一类数据对应的标签全是好或是坏，这时候就需要把1合并掉  
 如果同时都包含，不用合并。  
 如果var>5时，直接进行卡方分箱，默认先分为5箱，分完之后检查每个变量的每一箱是否满足bad rate就是好坏都存在，  
 如果不满足bad rate  
 继续分箱，分箱个数变为4个，然后检查bad rate，知道每箱bad rate 满足条件，当只剩两个分箱的时候必然是单调的  
   
 2. WOE编码，VI值计算  
 将每一个分箱里面的每一个变量的数据转换为WOE值，第i个分箱中该变量的某一个数据 WOE = god(i)/god(sum)/bad(i)/bad(sum)， 分子分母都不能为0，这就是为什么要保证每一个分箱的单调性  
 WOE其实是表征了某一个分箱的某一个变量两类样本分布的差异性  
 比如 某个分箱某个变量： good；0.8 good；0.2 good；0.1  
 bad 0.2 bad 0.2 bad 0.2  
 WOE 4 1 0.5  
 VI值是特征信息读，表征了这个分箱的区分能力 good - bad   
 作用: 用来挑选变量  
 1 高VI值表征了该特征与目标变量关联度高  
 2 目标变量只能是二分类  
 3 过高的VI值有潜在的风险  
 4 0.02-0.1 预测性弱 0.1-0.2 一定的预测性 0.2+ 高预测性

不同的特征在不同的分箱呈现的bad rate不同，比如一个好的模型分数，它的分箱肯定应该是严格单调的，而且不同分箱的bad rate差别还比较大，而比如年龄这样一个变量的分箱应该是不单调的，具体还是要看特征的含义

看一个具体的例子，两个三方分数对比：

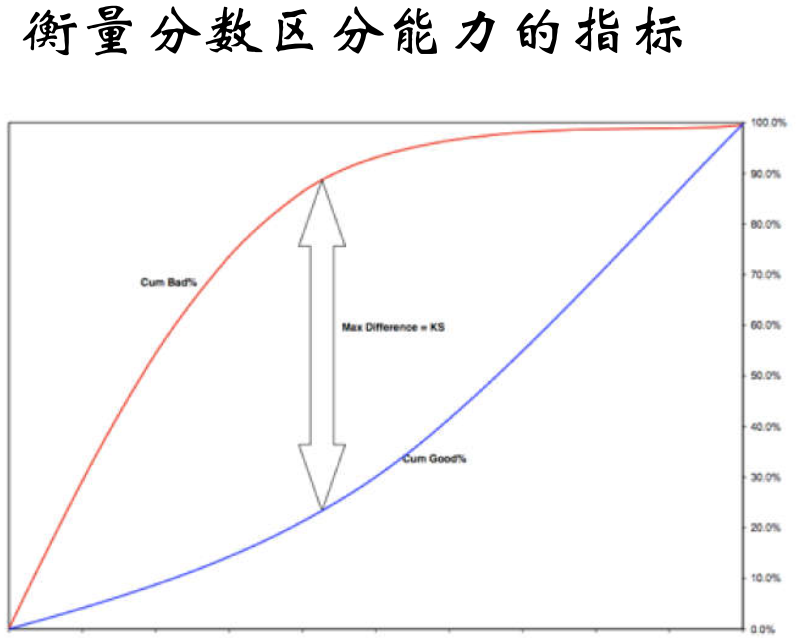
龙井分分箱：

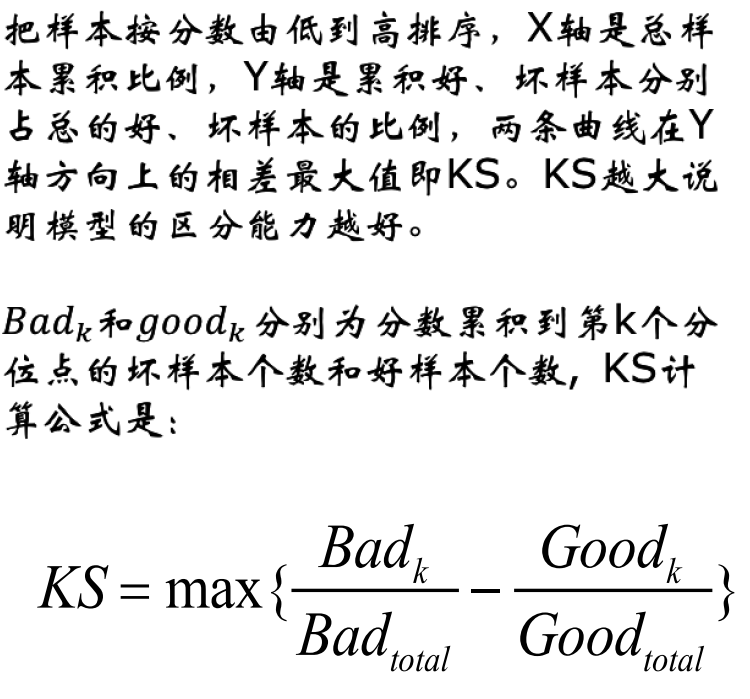


蚁盾分分箱：

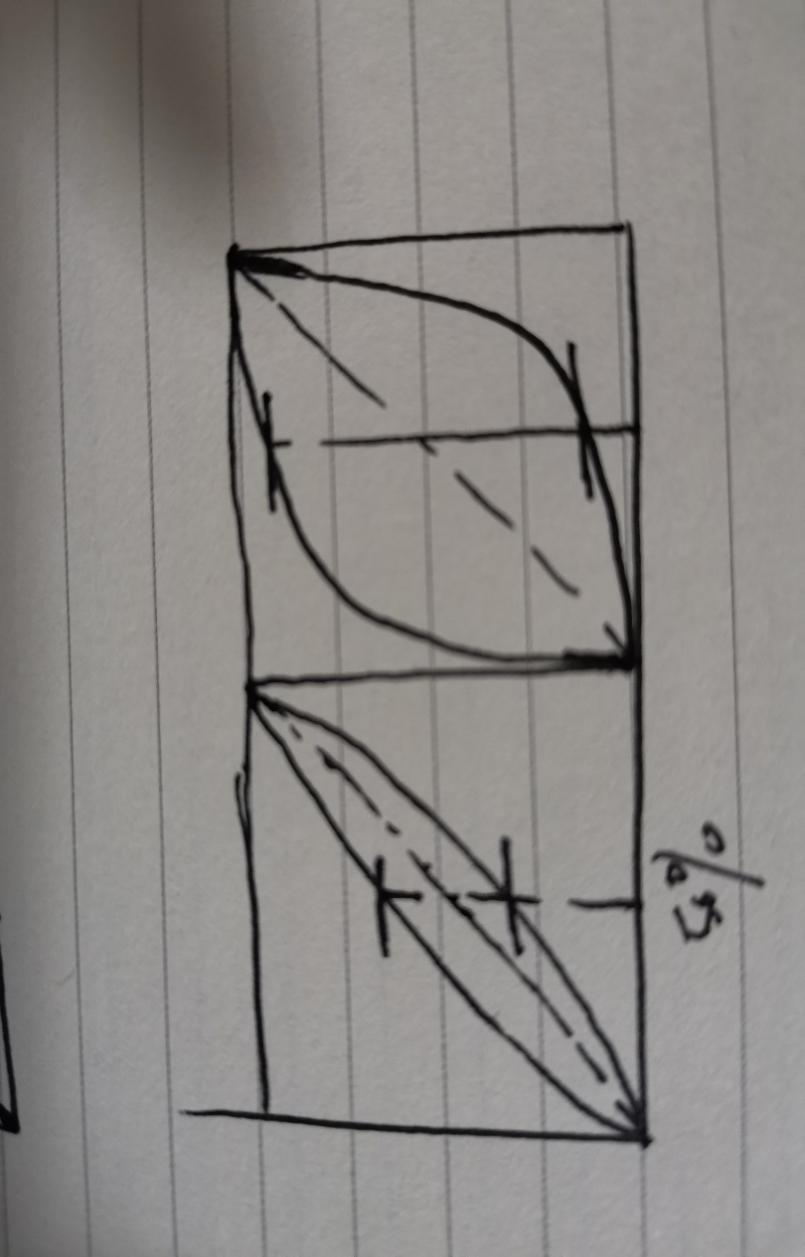
#### image2020-1-6_16-21-33 龙井分分箱结果也比较合理，单调性更好

验证模型   
 KS： 模型区分度  
 理解：正负样本区分度，好人坏人区分度，利用KS曲线来理解 随着样本的累计，模型预测的好样本占比（good/good sum）与坏样本占比(bad/bad sum)  
 之间的差值表征了模型的区分能力，max(good/good sum - bad/bad sum)就是模型的ks值表征了模型的区分能力。





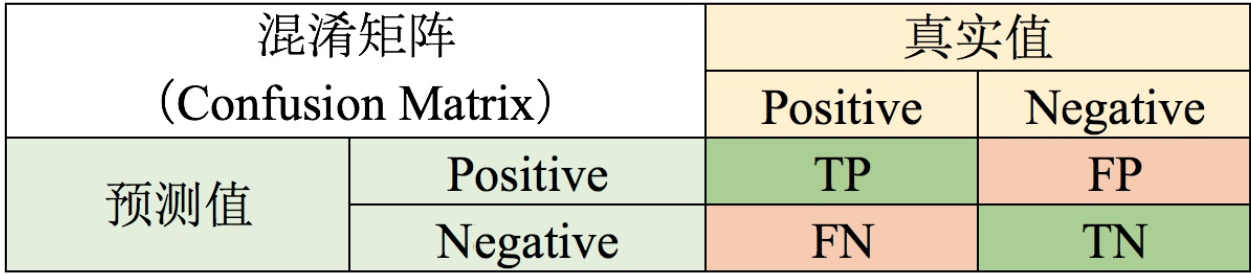
这里关键是要知道X轴的定义，把样本按分数从低到高排序，就是把模型预测样本的概率值从高到低排序，得到有序的预测结果序列，X轴表征样本累计占比。试想有一批样本好坏比2:8 当样本累计预测到了50%，这50%都是概率值比较高的样本，然后我们去看Y轴 得到了这50%的样本中goodrate和badrate。一个区分能力较弱的模型，虽然都是概率值比较高的样本 但是预测的结果中，goodrate，badrate很接近，很明显的区分能力弱，如果区分能力强的话，这50%的样本中bad/bad sum 应该非常高，而good/good sum 又很低 因为这些样本均概率值是很高的，那么bad/bad sum good/good sum 离得越远，区分能力当然就越好。随着样本累计100%，bad/bad sum good/good sum 都接近1了，KS曲线可以很好的表征模型区分能力



阀值： >0.3 好用 0.2-0.3 可用 0-0.2 较差 < 0 模型有错误

混淆矩阵的理解

在真正的风业建模过程中，样本不均衡问题是普遍存在的，正负样本比一般在1:20甚至更低，如果使用模型的准确率来评估模型是不可靠的，例如100个样本中好样本95个，坏样本5个，主要模型全部预测为好样本，模型的准确率就高达95%，因此业内更倾向于使用混淆矩阵来评估模型的效果。



混淆矩阵

前一位代表模型预测是否准确，True or False，后一位代表预测结果正负，Positive or Negative，通过组合可以代表所有样本的预测情况。

这个矩阵不好理解，我们转换一下

|  |  |  |
| --- | --- | --- |
|  | 真实值 Positive | 真实值 Negative |
| 预测 Positive | 1 1 | 1. 0 |
| 预测 Negative | 0 1 | 0 0 |

T F 是预测结果与真实结果比较结果 P N 模型预测结果

真阳率：TPR = TP / (TP + FN) 即模型预测准确的正样本在所有正样本中的占比

假阳率：FPR = FP / (FP + TN) 即模型预测错误的正样本在所有负样本中的占比

TRP、FRP用户描述模型对于正样本的预测情况



TP TN 是预测对的

精准率(precision) TP / (TP + FP) 查准率 (TP + FP) 代表所有预测的正样本

精确率是针对我们预测结果而言的，它表示的是预测为正的样本中有多少是真正的正样本 更加关注 模型对于真正的正样本的预测能力

召回率(recall) TP / (TP + FN)

FN 代表预测为负样本标签预测错了，代表真实正样本 (TP + FN) 代表所有真正的正样本

召回率其实是看在所有模型预测的正样本中 准确的是多少，召回了多少

F1-score(均衡平均数) 2TP / (2TP + FP + FN)

混合的度量，对不平衡类别非常有效

是综合考虑了模型查准率和查全率的计算结果，取值更偏向于取值较小的那个指标。

F1-score越大自然说明模型质量更高。但是还要考虑模型的泛化能力，F1-score过高但不能造成过拟合，

影响模型的泛化能力

准确率(accuracy) TP + TN / (TP + FP + TN + FN)

模型的整体的性能的评估

为什么要出现这么多评估指标呢？实际上，不同的分类任务适合使用不同的指标来衡量。

例如，推荐系统中，如果希望更精准的了解客户需求，避免推送用户不感兴趣的内容，因为推送的内容都是模型预测的正样本，就是TP+FP,故precision 就更加重要；

在疾病检测的时候，我们不希望查漏任何一项疾病，这时 recall（TPR） 就更重要。

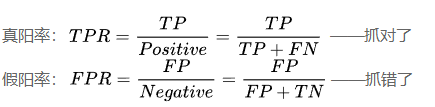
当两者都需要考虑时，F1-score 就是一种参考指标。

真实世界中的数据经常会面临 class imbalance 问题，即正负样本比例失衡，

而且测试数据中的正负样本的分布也可能随着时间变化。根据计算公式可以推知，

在测试数据出现imbalance 时 ROC 曲线能基本保持不变

AUC： 模型准确度



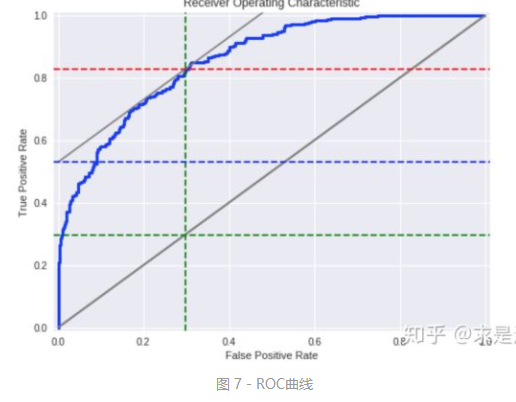
为简化考虑，我们的风控系统只有一个信用评分模型，那么高于分数阈值就预测为good，反之为bad，予以拒绝。我们一般追求更高的TPR，也就是"抓对了"；以及更低的FPR，也就是"抓错了"。抓对了和抓错了——是在模型决策的一念之间（不同的分数阈值），因此我们会设置不同的阈值来观察这种变化规律。

接下来介绍ROC曲线的绘制方法（后面将会以ROC曲线来辅助我们理解KS曲线）。

step 1. 先按分数升序排列，计算某个阈值T下分数区间[0，T]内的TPR和FPR。可以把TPR理解为累积正样本率(cum\_bad\_rate)，FPR理解为累积负样本率(cum\_good\_rate)

step 2. 重复step1多次，在不同阈值T下计算得到多个TPR和FPR。

step 3. 以FPR为横轴，TPR为纵轴，画出ROC曲线（如图7）。曲线下方的面积即为AUC值。

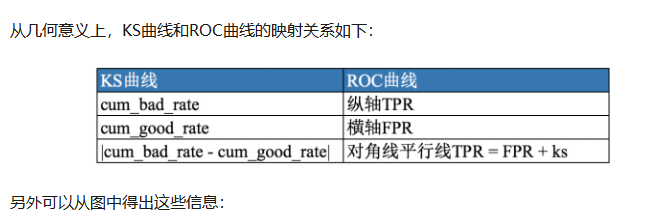


我们该如何理解ROC曲线？

如果模型没有任何排序性，那么正样本在每个分数区间上均匀分布。此时，无论如何设置阈值，[0，阈值]这个分数区间内TPR将与区间长度成正比关系，所以就是对角线。

如果希望TPR尽可能高，FPR尽可能低，我们可以设计一个目标函数，这与KS的定义完全相同。

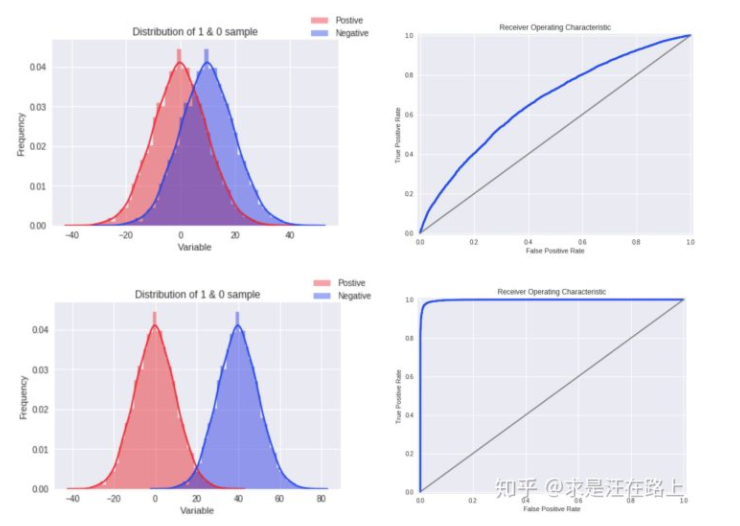




若希望KS尽可能大，那么R需要尽可能接近（0，1），此时AUC一般也会增大。对于相同的KS值，在KS曲线上有两个选择，但TPR和FPR同时大或同时小。虽然我们的目的通常是抓对更多的坏人（TPR⬆），尽可能减少错抓的好人（FPR⬇），但两者需要trade-off。到底选择哪个阈值，取决于业务目标：是希望对bad有更高的召回，还是对good有更低的误伤？

由于KS只是在一个最大分隔点时的值，并不够全面。通常我们也会同时参考KS和AUC

在理解KS和ROC曲线的关系后，我们也就更容易理解——为什么通常认为KS在高于75%时就不可靠？我们可以想象，如果KS达到80%以上，此时ROC曲线就会变得很畸形，如图9所示。



另一个更重要的可能原因是，为了便于制定策略，模型评分在放贷样本上一般要求服从正态分布。如果出现这种明显的双峰分布，就不太符合业务sense

阀值： >0.7 区分度高 0.6-0.7 有一定的区分度 0.5-0.6 区分度较弱 <0.5 基本和瞎猜没啥区别