信用评分

张兴敏

西南财经大学

Southwestern University of Finance and Economics

第四章 特征工程

风控模型开发流程

- 数据获取：从各数据源获取原始数据
- 数据处理：对原始数据进行整理变换
- 特征工程：确定有预测能力、符合逻辑、稳定可得、合规、有关、相关性低的特征作为备选变量
- 样本细分：根据市场、渠道、客户、数据、流程等因素对样本进行分层，分别建立评分卡
- 训练模型：用历史样本建立模型
- 拒绝推断：开发申请评分卡时对被拒绝的申请者的表现进行推断。
- 模型校准：确保分数在不同评分卡中有相同的含义，反映相应的违约概率。
- 验证交付：用保留样本和近期样本检验模型是否过度拟合或不稳定，然后准备投入使用。

数据准备

数据来源：客户、内部、外部

- 申请表格——采集并编译，现多为电子表格，或不需表格。
- 征信机构——查询、回溯检索。
- 内部系统——从内部系统获取行为数据。
- 表现数据——每个样本的好坏结果。
- 匹配合并——将所有数据表整合到一起。匹配键key？

数据准备

数据校验data validation：对数据的有效性进行查验

- 例如，没有150岁的人，没有21岁就已持有银行账户超过30年的人。
- 发现有特殊情况的，不应被简单理解和处理，而应用特殊代码如9999标记。如果存在缺失数据，最规范做法是把它们作为缺失值进行编码，而不是试图估算填入一个实值。
- 实际操作中，缺失数量不太多时，也可以替代为某个实值（平均值）
- 真实数据包含大量的缺失值、极端值、异常值

样本设计

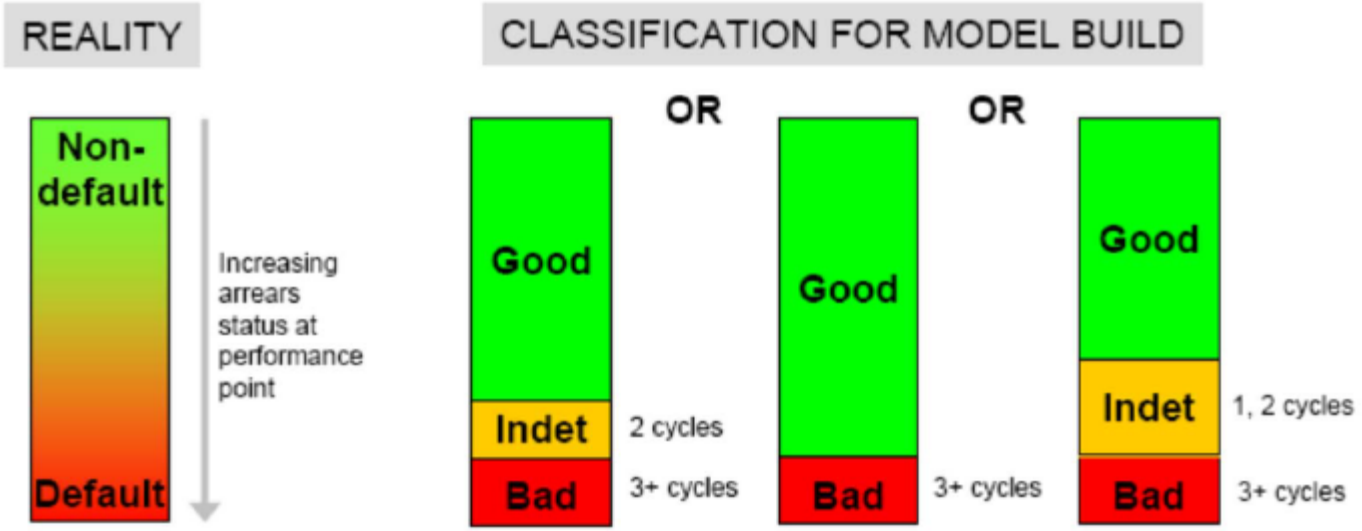
好坏定义

坏账户：default/charged off，90天/连续三期逾期（M3）——巴塞尔协议

好账户：Fully paid (prepayment)

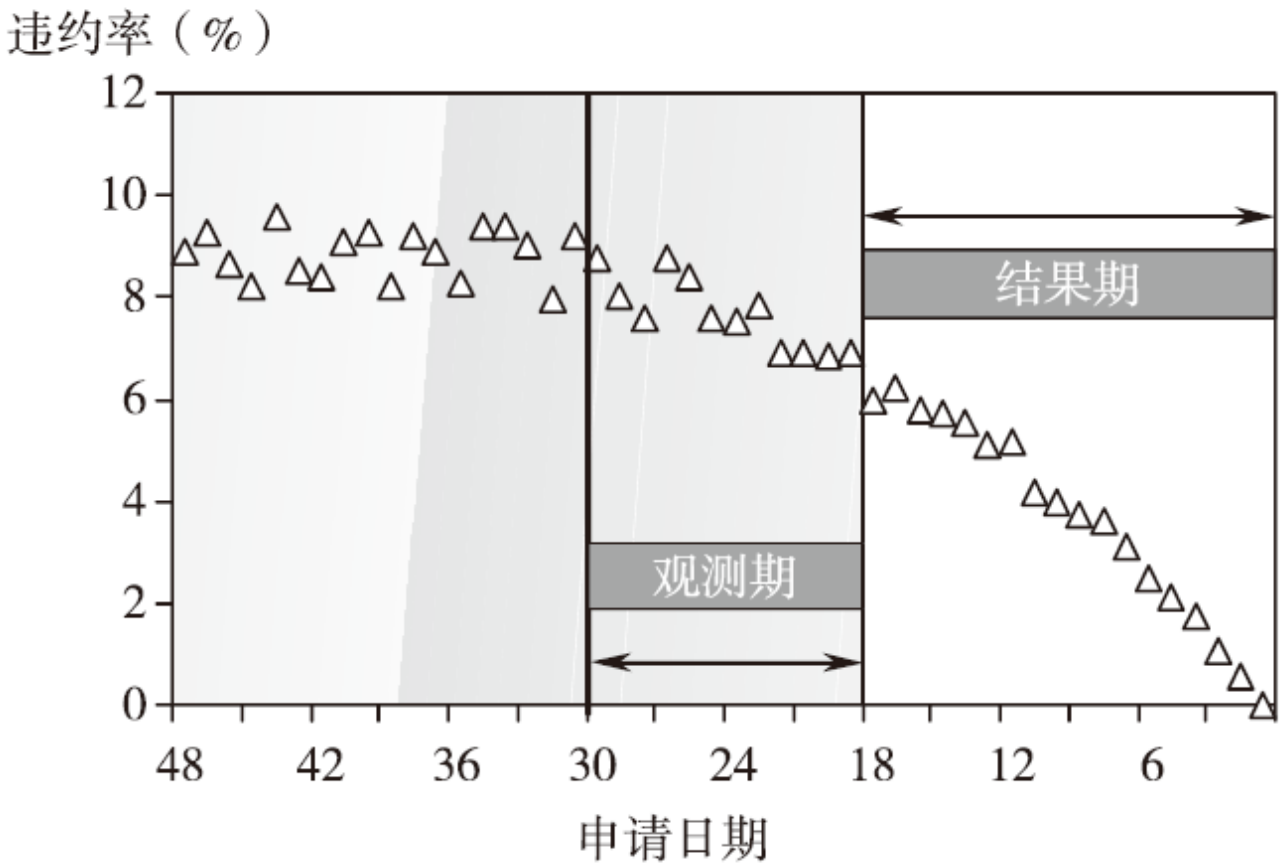
正常还款账户：Current; up-to-date

不确定的账户：M1 or M2



样本设计

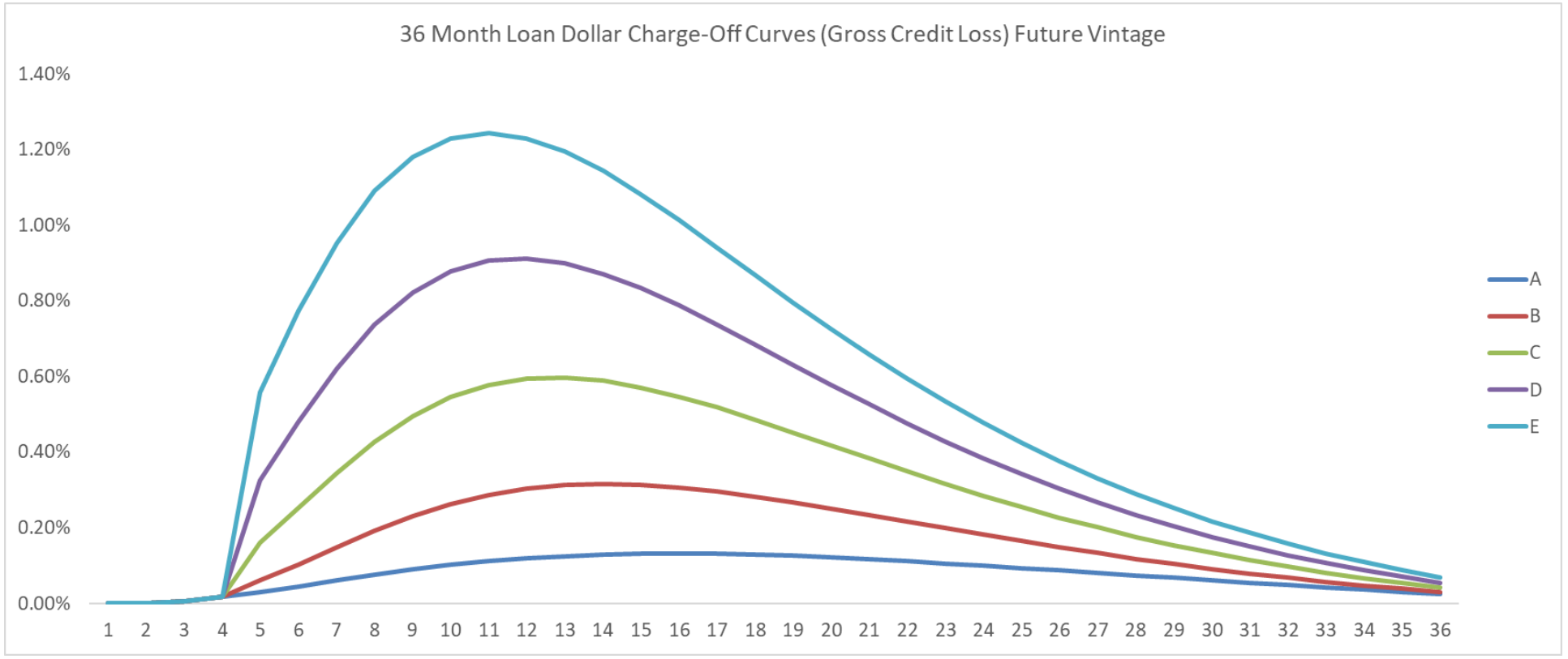
违约成熟



样本设计

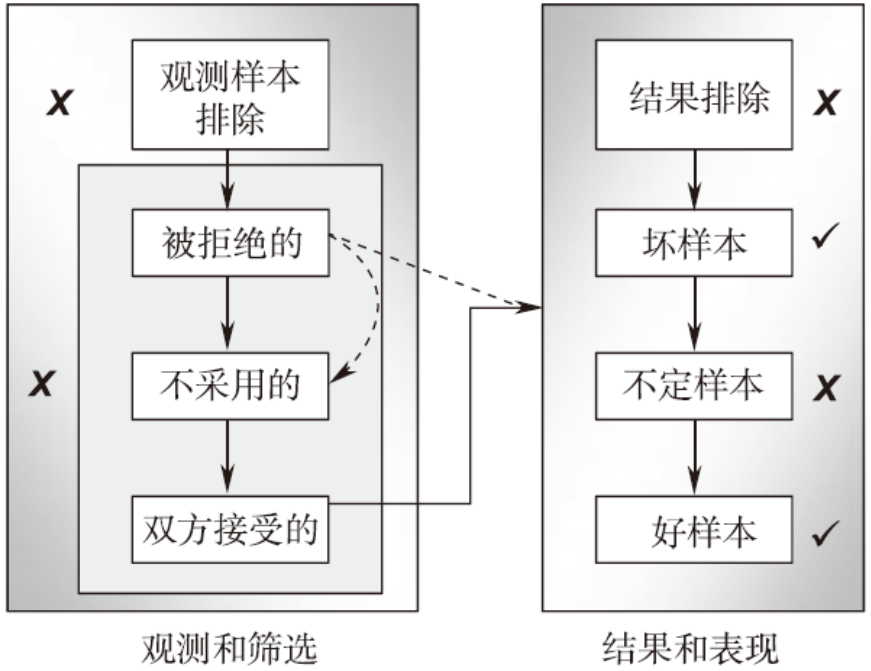
账龄分析vintage analysis

Month on Book



样本设计

- 训练样本training sample——用来开发预测模型的样本，这是最关键的部分，需要大量的观测。
- 保留样本（测试样本）hold-out or test sample——历史样本中单独保留的样本，用来验证模型，模型可能过度拟合，在训练样本表现很好但在其他样本上没用。



样本设计

- 样本分层segmentation决定是否要细分总体并为每部分单独建模。
- 例如，为25岁以下和25岁及以上的人各建立一个评分卡，或者为那些高收入的人建立一个评分卡，也为低收入的人建立另一个。
- 建立多个评分卡会带来大量额外的工作，所以只在理由充分并且改进预测效果时使用。
- 样本分层的理由有如下几个：

  数据可得性、特征相关性、经营策略

- 通用评分卡generic model和分层评分卡segmented model：一般通用模型普适性好，专用模型预测性好。

特征工程

- 变量剔除：稳健的模型通常有10到20个特征变量，而可用的特征远远多于这个数量，需要剔除变量。
- 剔除的特征：区分好坏的能力偏小；与其他确定要使用的变量高度相关甚至存在共线性；在时间上不稳定。
- 多重共线性（multicollinearity）：严重危害模型！

  变量不显著；符号不确定；显著性不稳定；过度拟合
- 多重共线性解决办法：相关性分析、变量剔除、逐步回归、主成分分析

特征工程

	Own phone	No phone	
Owner	95%	50%	90%
Tenant	75%	65%	70%
	91%	60%	

- 自有房产的人比租房的人的好人比例更好（90% vs 70%），有电话的人比没电话的人的好人比例更高（91% vs 60%），
- 在线性模型里，有电话有房产的人的得分最高，得分最低的是没有电话的租房者。事实上，有房产但没有电话的人实际上更坏（50%）。
- 如果事先不知道这个现象，这种非线性关系很难被线性模型捕捉到。
- 决策树、神经网络、支持向量机、贝叶斯网络都能很好的应对这种非线性关系。建模人员可以先用决策树发现变量的相互作用，然后依次进行样本分层建模。
- 在这个例子里就是为自有房产和租房的人单独建模。一般年龄是主要的分层变量。

方法一：朴素贝叶斯 (naïve Bayes)

- **基本思想：**基于先验信息（如好坏分布）和样本信息（如申请者的特征）推断后验分布（如对数比率分数等）
- 朴素贝叶斯（naïve Bayes）评分卡: 定义一个**对数比率分数（log odds score）为信用分数**
- **假设：**评分卡的特征相互独立；

- **信贷审批的决策：**通过发放贷款或拒绝，“好人”或“坏人”两个类别；
- 设$\mathbf{X} = (X_1, X_2, \ldots, X_n)$ 是借款人特征，如年龄、婚姻、住房等；
- 设$\mathbf{x} = (x_1, x_2, \ldots, x_n)$ 是借款人特征的属性值，如年龄的属性有：18-25岁，26-35岁，36-43岁,>43岁等；
- p(G) 和 p(B) 是先验概率；

方法一：朴素贝叶斯

- 后验概率 p(G|**x**) 是给定某些属性值时借款人是**好人**的概率

  p(G|**x**) =（申请人具有属性**x**且是好人的概率）/（申请人具有属性**x**的概率）

- 后验概率 p(B|**x**) 是给定某些属性值时借款人是**坏人**的概率

  p(B|x) =（申请人具有属性x且是坏人的概率）/（申请人具有属性x的概率）

- p(x|G) 和 p(x|B) 是在好人或坏人总体中，属性值**x**的似然值
-
- 例如：

	好人	P(x|G)	坏人	P(x|B)	好人比率
已婚	4900	0.7	400	0.4	4900:400=12.25:1
未婚	2100	0.3	600	0.6	2100:600=3.5:1
合计	7000	1	1000	1	

朴素贝叶斯

朴素贝叶斯（naïve Bayes）评分卡

定义一个对数比率分数（log odds score）为信用分数 $s(\mathbf{x})$

朴素贝叶斯方法假定特征之间相互独立

$$p(\mathbf{x}|G) = p(x_1|G)\,p(x_2|G)\cdots p(x_n|G)$$

$$p(\mathbf{x}|B) = p(x_1|B)\,p(x_2|B)\cdots p(x_n|B)$$

对数比率评分卡

$$s(\mathbf{x}) = \ln\left(\frac{p(G|\mathbf{x})}{p(B|\mathbf{x})}\right) = \ln\left(\frac{p_G\,p(\mathbf{x}|G)}{p_B\,p(\mathbf{x}|B)}\right)$$

$$= \ln\left(\frac{p_G}{p_B}\right) + \ln\left(\frac{p(x_1|G)}{p(x_1|B)}\right) + \ln\left(\frac{p(x_2|G)}{p(x_2|B)}\right) + \cdots + \ln\left(\frac{p(x_n|G)}{p(x_n|B)}\right)$$

$$= \ln(o_{pop}) + \mathrm{woe}(x_1) + \mathrm{woe}(x_2) + \cdots + \mathrm{woe}(x_n)$$

例子：朴素贝叶斯

表 3.1　朴素贝叶斯的例子

	有房		无房		合计	
年龄	好人	坏人	好人	坏人	好人	坏人
30 –	100	10	200	40	300	50
30 +	500	10	100	40	600	50
合计	600	20	300	80	900	100

朴素贝叶斯评分卡s(**x**)？

一个35岁，有房子的人的信用分数是多少？

评分卡建模流程

01	02	03	04	05
数据准备	训练模型	加总点数	筛选	决策
足量历史样本（好坏样本）； 借款人申请信息和结果表现label（B0/G1）	找到特征变量和属性对应的权重（系数coefficient或点数point）	得到借款人信用分数score	设定合理临界值/阈值cut-off	作出决策decision

分析建模

- 统计方法

线性回归，判别分析，**逻辑回归**，分类决策树，生存分析……

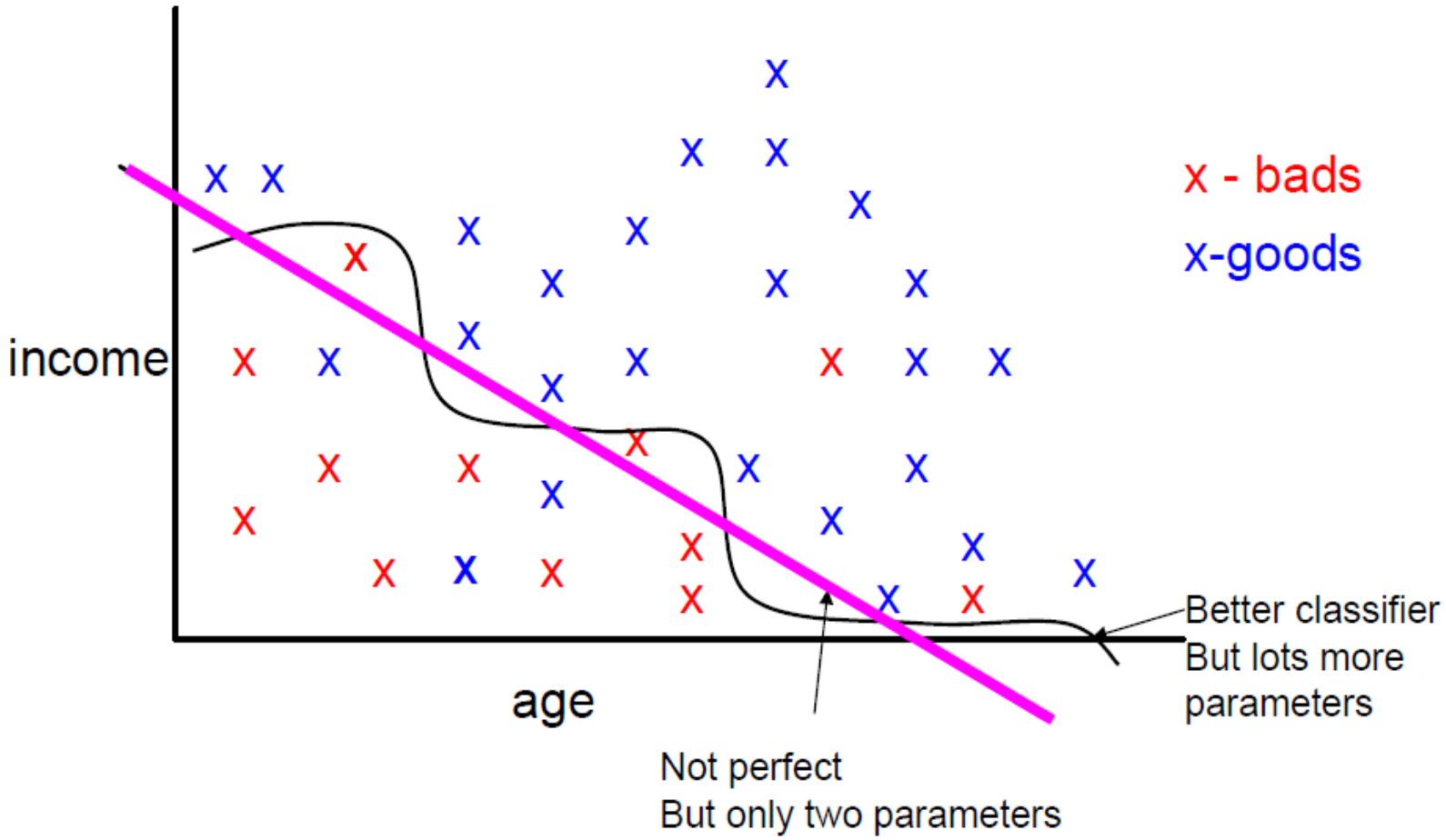
- 非统计方法

线性规划、遗传算法、神经网络、**随机森林**、支持向量机……

- 不论用什么方法，最后得到的结果是根据每个个体特征计算出来的一个具体的**数值（分数）**，然后以此进行好坏判定。
- 建模之前：变量的描述统计分析很重要！

第五章 模型评价

预测好坏


x - bads
x-goods
income
age
Better classifier
But lots more
parameters
Not perfect
But only two parameters

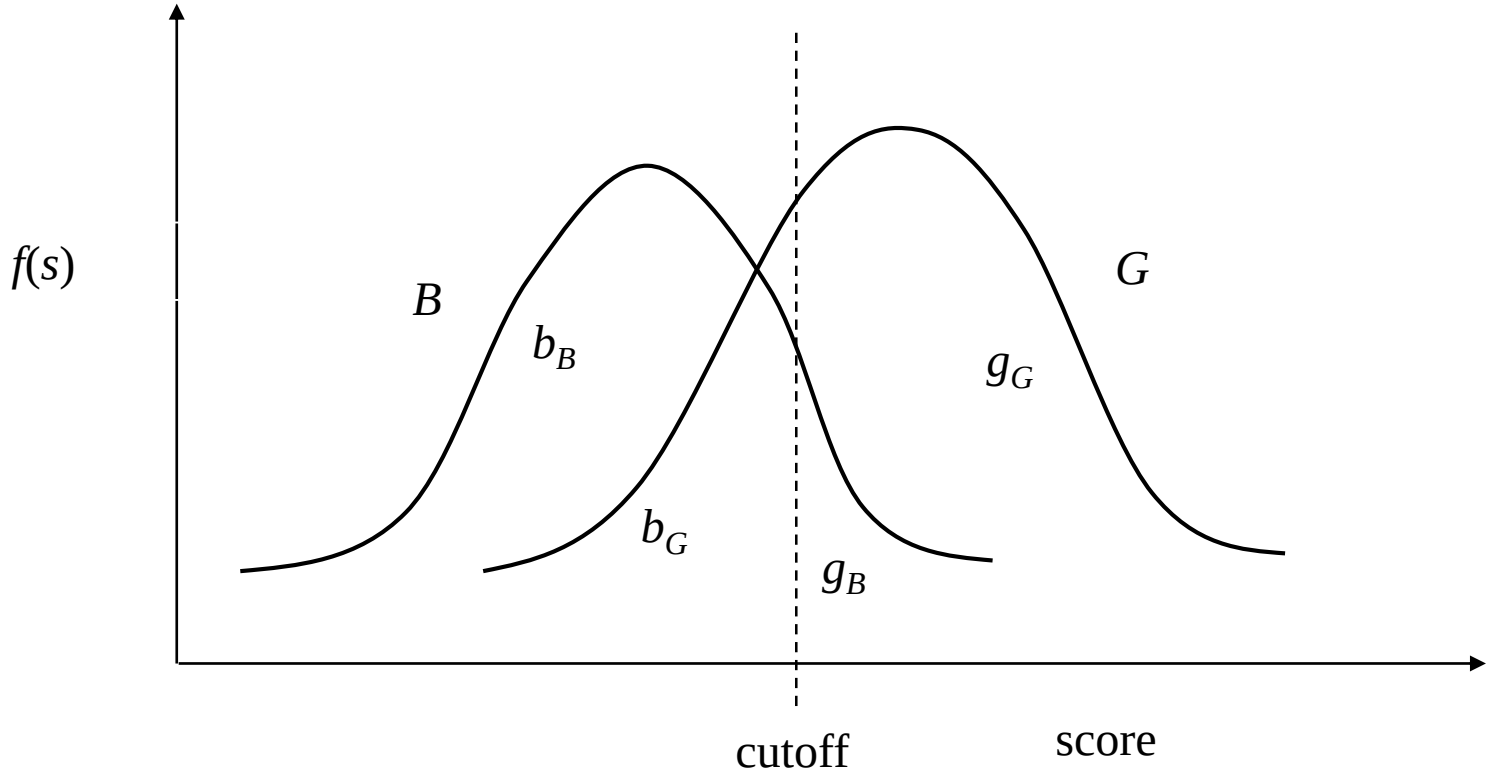
预测好坏

- 将好人分数排序（分数越高，越可能是好人），或者将违约概率排序（违约1，否则0，越接近0，越可能是好人）
- 选定一个cut-off（阈值、临界值、分界线、合格线）

	……	556	558	573	575	580	589	590	600	620	……
真实		坏	坏	好	好	坏	好	坏	坏	好	
预测		坏	坏	坏	坏	坏	好	好	好	好	

预测好坏

- 分数分布



预测好坏

- 混淆矩阵

		真实		预测
		好人	坏人	
预测	好人	正确好人	二类错误	预测好人
	坏人	一类错误	正确坏人	预测坏人
真实		真实好人	真实坏人	总数

第一类错误Type I error：原假设为真（好人），但拒绝了原假设（坏人）

第二类错误Type II error ：原假设不真（坏人），但没有拒绝原假设（好人）

直观上来讲：第一类错误是把好人预测成坏人（左下单元格）；第二类错误是把坏人预测成好人（右上单元格）

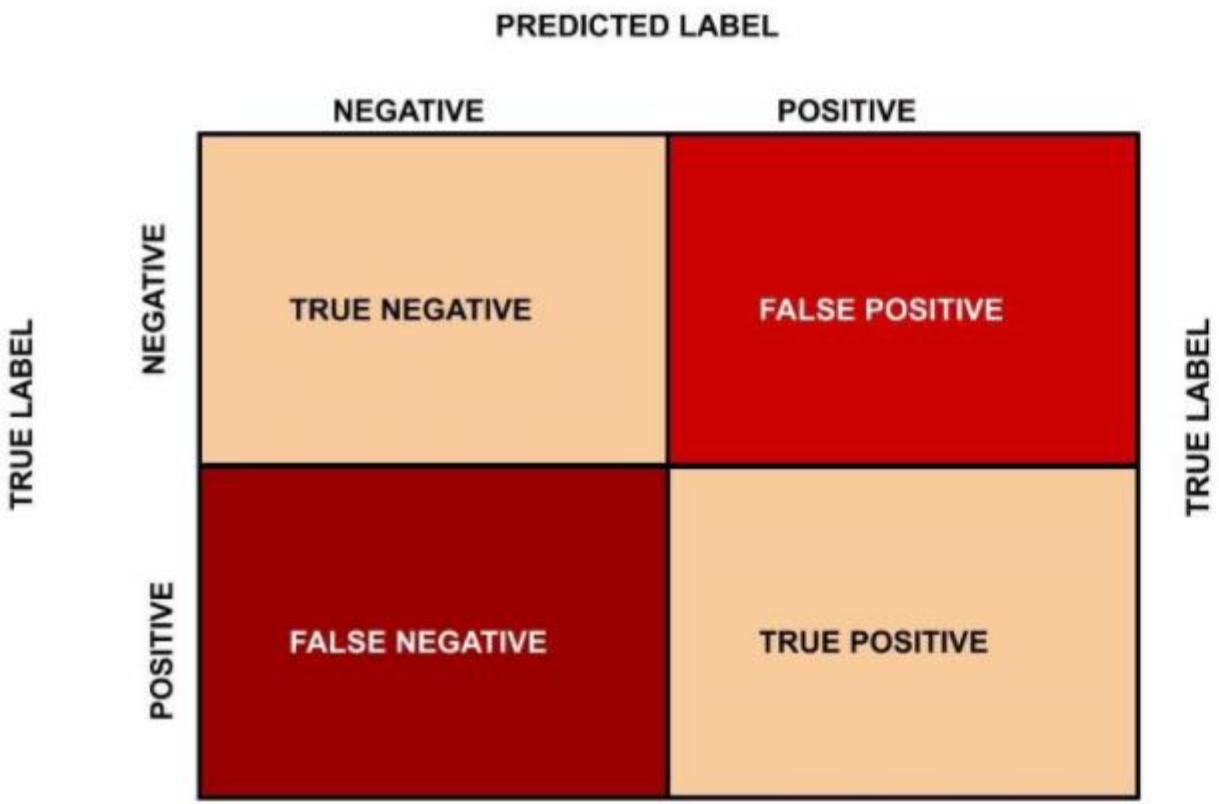
预测好坏

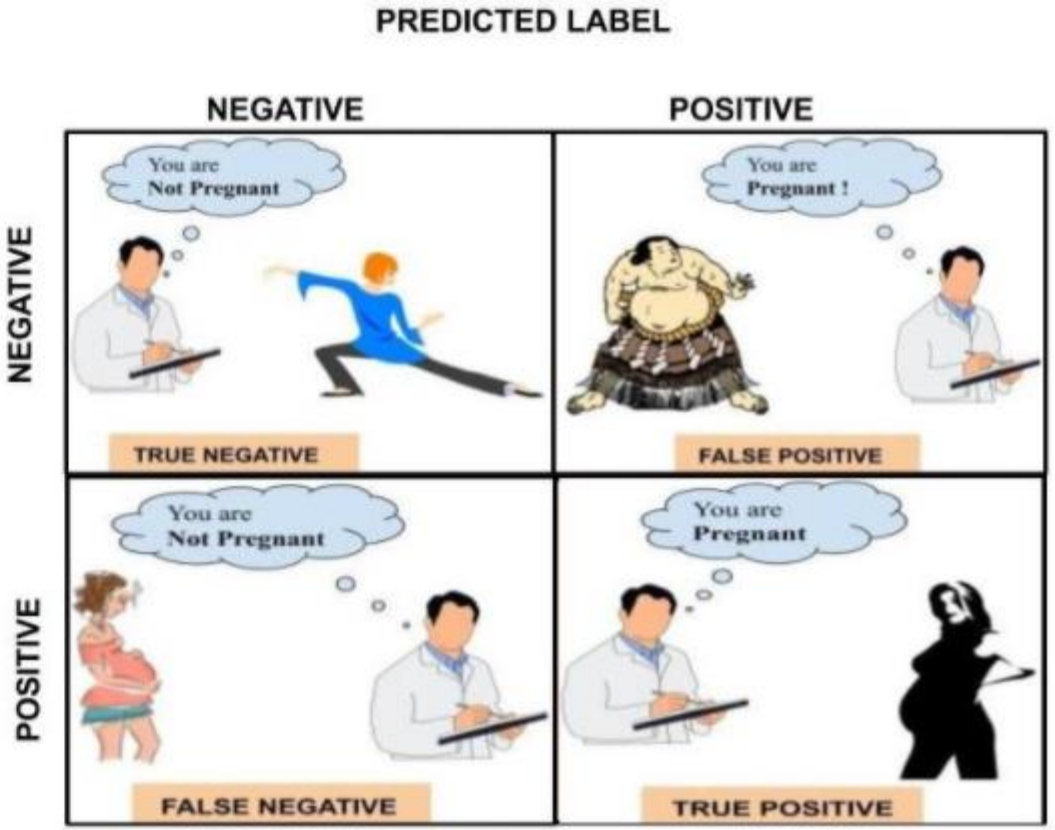
- 混淆矩阵

错误类型		空假设H0	
		真True	假False
关于空假设的决策	无法拒绝	正确推断（真阴性）（概率=1-α）	二类错误（假阴性）（概率=β）
	拒绝	一类错误（假阳性）（概率=α）	正确推断（真阳性）（概率=1-β）

- 第一类错误（type I error rate）或显著性水平（significance level）（α level）：在给定为真的情况下，拒绝原假设（空假设）的概率
- 第二类错误（type II error rate）记为β (beta)，与功效（power=1- β）
- 空假设："is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis. **"— Fisher, 1935, p.19**

预测好坏


PREDICTED LABEL
NEGATIVE
POSITIVE
TRUE LABEL
NEGATIVE
POSITIVE
TRUE NEGATIVE
FALSE POSITIVE
FALSE NEGATIVE
TRUE POSITIVE
TRUE LABEL


PREDICTED LABEL
NEGATIVE
POSITIVE
NEGATIVE
POSITIVE
You are Not Pregnant
You are Pregnant !
You are Not Pregnant
You are Pregnant
TRUE NEGATIVE
FALSE POSITIVE
FALSE NEGATIVE
TRUE POSITIVE

预测好坏

◆ **预测正确：**

预测好人占总体好人的比例 = 特异度 = g_G/n_G

预测坏人占总体坏人的比例 = 敏感度 = b_B/n_B

◆ **预测错误：**

预测坏人占总体好人的比例 = 第一类错误率 = b_G/n_G

预测好人占总体坏人的比例 = 第二类错误率 = g_B/n_B

	实际好人	实际坏人	预测人数
预测好人	g_G	g_B	g
预测坏人	b_G	b_B	b
实际人数	n_G	n_B	

特异度

敏感度

- 特异度specificity：真阴率true negative rate, 召回率recall，查全率
- 灵敏度or敏感度sensitivity：真阳率true positive rate，精确率precision，查准率
- 总预测准确性accuracy：$(g_G + b_B)/n$

预测好坏

- 错误分类成本misclassification cost（潜在损失）不对称！
- 第一类错误是把好人判定为坏人，拒绝他们会损失好人带来的潜在利润L。
- 第二类错误是把坏人判定为好人，接受他们会带来违约的损失D。
- 成本D和L可能都是未知的！
- 我们希望一个评分模型能够给贷款机构带来最大的期望利润，如果不行，至少是能够带来最小的错误分类成本（财务损失）。很少有研究关注这两个目标，困难在于我们无法知道成本D和L 。
- 大部分实证研究都只能给两类错误同样的权重（不考虑D和L ），然后计算预测准确性。

预测好坏

- 分界线：299

	实际好人	实际坏人	预测人数
预测好人	13798	118	13916
预测坏人	4436	765	5201
实际人数	18134	883	19117

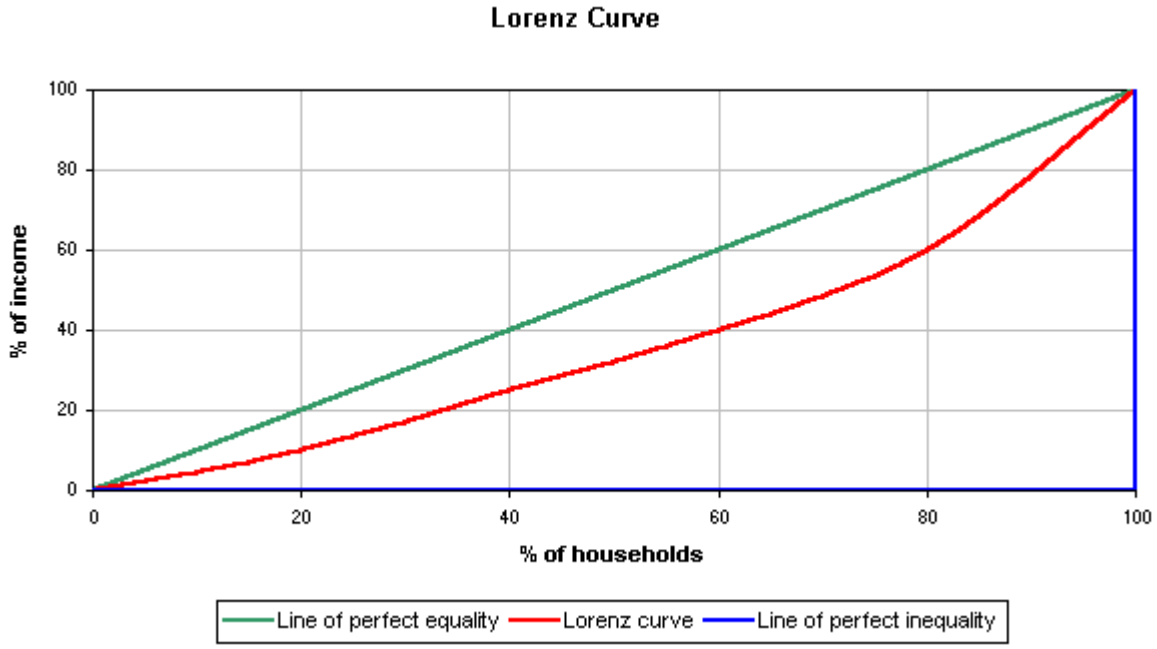
- 分界线：251

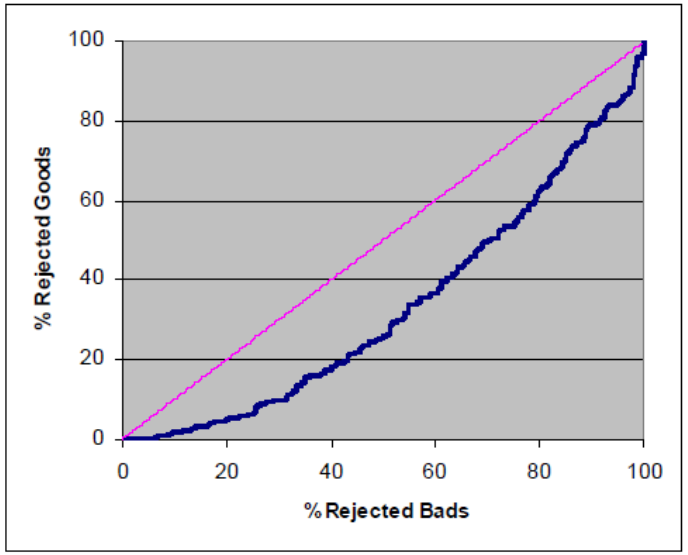
	实际好人	实际坏人	预测人数
预测好人	17282	424	17706
预测坏人	952	459	1411
实际人数	18234	883	19117

- 不同的临界分数线带来不同的准确性，如何综合评价？

模型效果：ROC

Lorenz曲线：1905年由经济学家洛伦兹提出用来表示收入分配的曲线，意大利经济学家基尼Gini在此基础上定义了基尼系数。





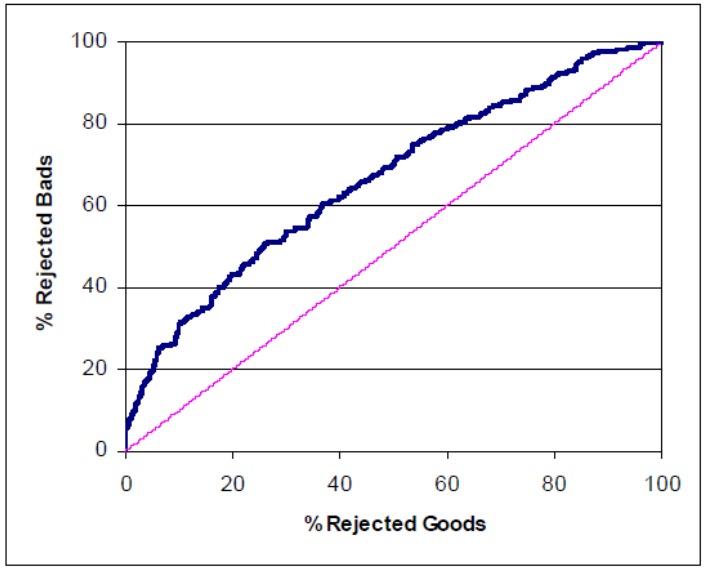
模型效果：ROC

接收者（受试者）操作特征曲线（receiver operating characteristic curve，ROC曲线）是一种座标图式的分析工具。

在做决策时，ROC分析能不受成本 / 效益的影响，给出客观中立的建议。

ROC曲线首先由二战中的电子工程师和雷达工程师发明，用信号检测理论来侦测战场上的敌军。之后很快就被引入了心理学来进行信号的知觉检测。几十年来，ROC曲线被用于医学、无线电、生物学、心理学领域中，最近在机器学习和数据挖掘领域也得到了很好的应用。

$F_b(s)$

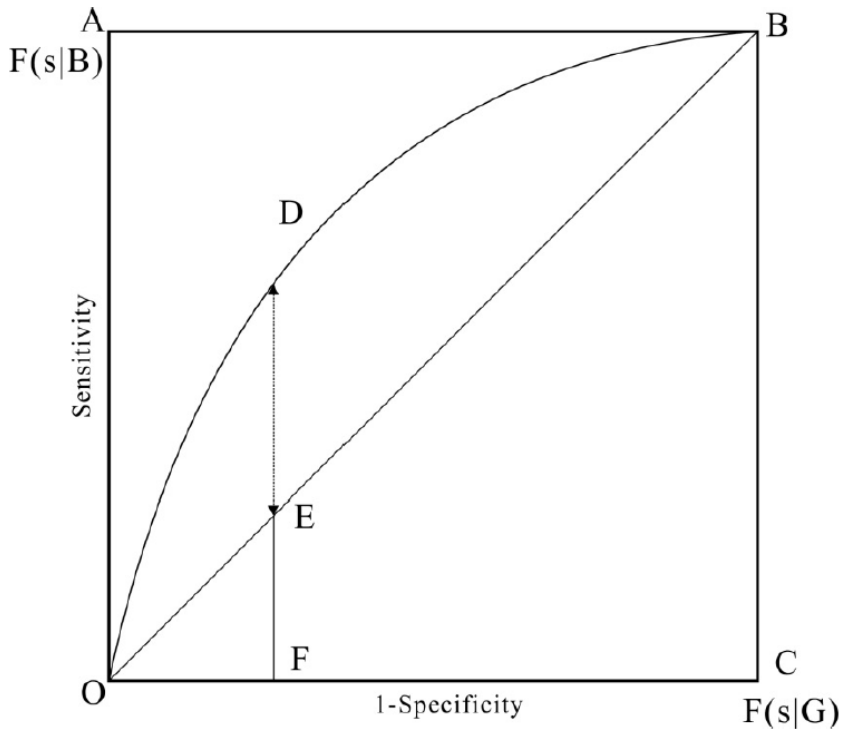


$F_g(s)$

模型效果：ROC

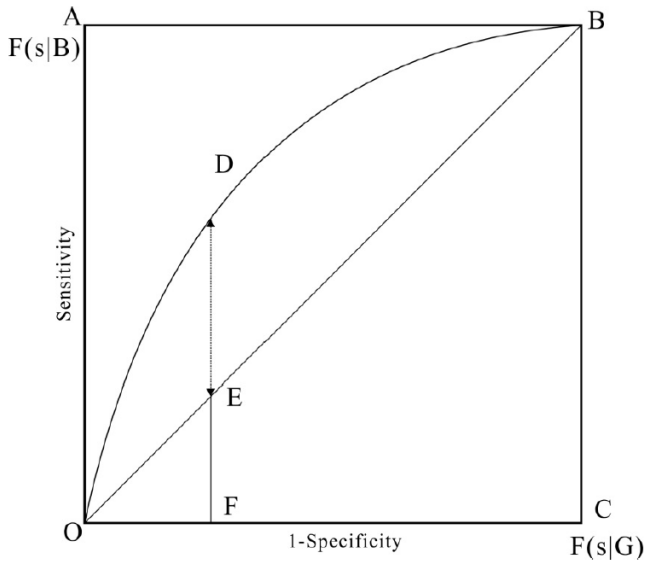
ROC曲线

- 横坐标：（1-特异度） F(s|G) ，在分数s下的好人累积比例
- 纵坐标：敏感度 F(s|B) ，在分数s下的坏人累积比例
- 总体表现用曲线下方的面积AUC（Area Under the ROC Curve）来评价。

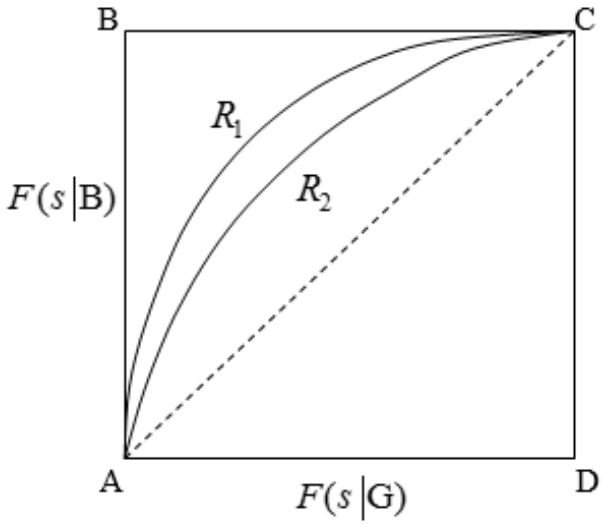


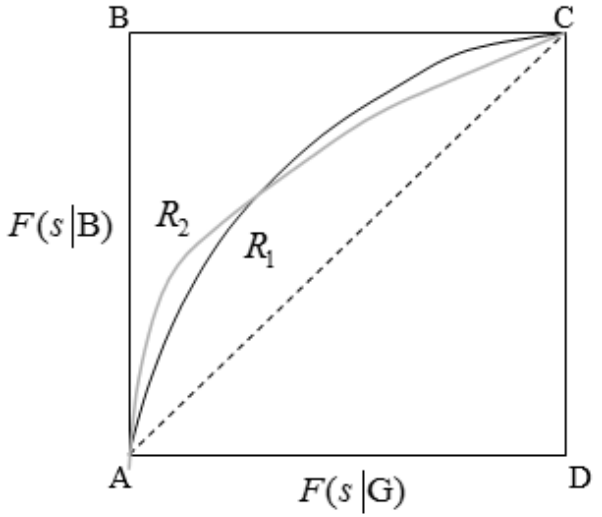
模型效果：ROC

- 完美评分卡：OAB
- 随机评分卡：OEB
- 一般评分卡：ODB
- ROC曲线越接近直角边，评分卡表现越好；ROC曲线越接近对角线，评分卡表现越差；
- 完美评分卡的AUC是1，随机评分卡的AUC是0.5
- 当曲线在对角线下方时，可以将分类器逆向使用
- 这样，AUC的实际值分布在0.5~1之间，面积越大，判别能力discriminant power越好



模型效果：ROC



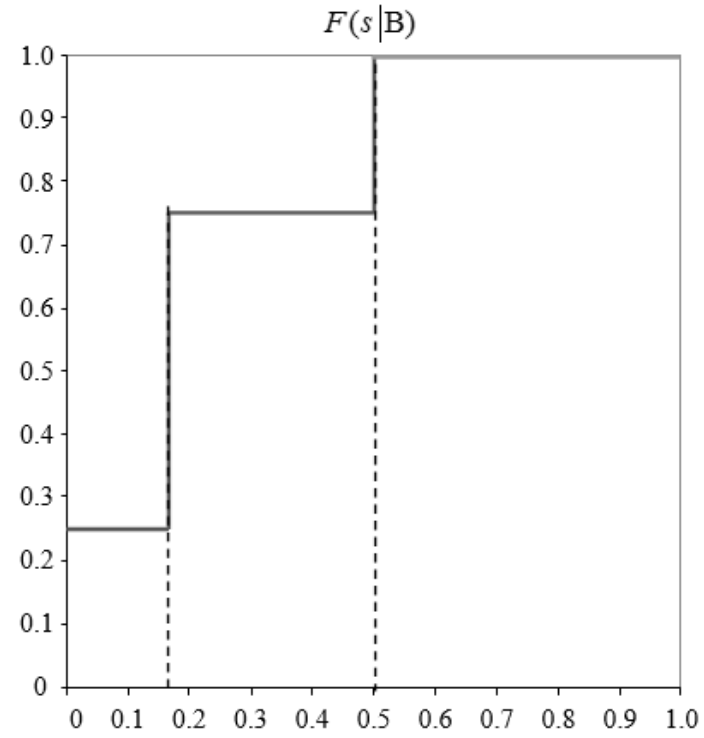


- 如果两条ROC曲线总有R1在 R2上方，则表明在所有分数上，第一个评分卡都比第二个评分卡能更好区分好坏。
- 如果两条曲线相交，其中一个评分卡在一个区域内更好但另一个评分卡在该区域外的地方更好。在低分区，R2 能更好区分好坏，但在大的临界分数上， R1更好。

模型效果：ROC

- 绘制ROC曲线

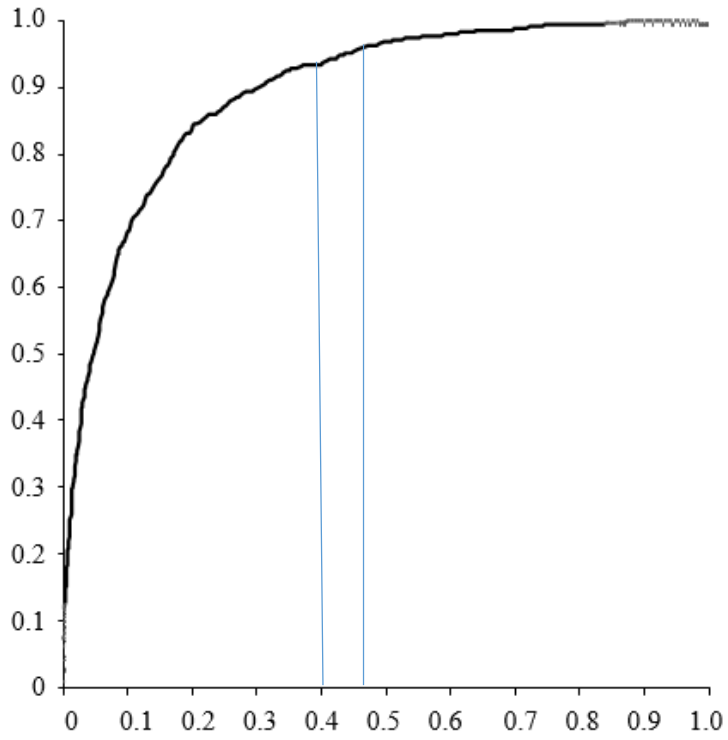
已知评分卡在 10 个借款人的样本上检验，其中有四个坏人（分数是 150、190、200 和 250），六个好人（分数是 180、205、230、260、280 和 300）。在画 ROC 曲线和计算 Gini 系数时，分值大小并不重要，真正重要的是这些人分数的相对排序。把分数按升序排列，得到 BGBBGGBGGG。指定一个比所有分数都低的分，比如 140，低于此分数的好人、坏人比例都是 0，所以 ROC 曲线会通过原点 (0,0)。假如取分数 160，它只比一个坏人的分数高，低于其他所有人，有 1/4 的坏人和 0 的好人低于这个值，所以点 (0,1/4) 也在 ROC 曲线上。继续增加设定的分数，使得每次只有更多的一个人落在设定的分数下方，我们发现 ROC 曲线会经过点 (1/6,1/4)（一好一坏低于这个分），然后是点 (1/6,2/4)、(1/6,3/4)、(2/6,3/4)、(3/6,3/4)、(3/6,4/4)、(3/6,4/4)、(4/6,4/4)、(5/6,4/4) 直到 (6/6,4/4)。这样便画出了图 2.3.3 中的 ROC 曲线。



模型效果：ROC

- 当样本量增加到几千上万时，ROC曲线看起来更像是连续的曲线。计算它的面积，可以将曲线分成很多小段，分别计算每小段的矩形面积（宽度*高度），再加总得到。

分数段	好人数量	坏人数量	好人%	坏人%	梯形面积
143~277	2614	667	0.143	0.755	$0.143(0+0.755)/2=0.054$
278~319	3760	150	0.206	0.925	$0.206(0.755+0.925)/2=0.173$
320~349	3938	45	0.216	0.976	$0.216(0.925+0.976)/2=0.2053$
350~382	3952	15	0.217	0.993	$0.217(0.976+0.993)/2=0.2134$
383+	3970	6	0.218	1	$0.218(0.993+1)/2=0.2170$
合计	18234	883	1	1	0.8631



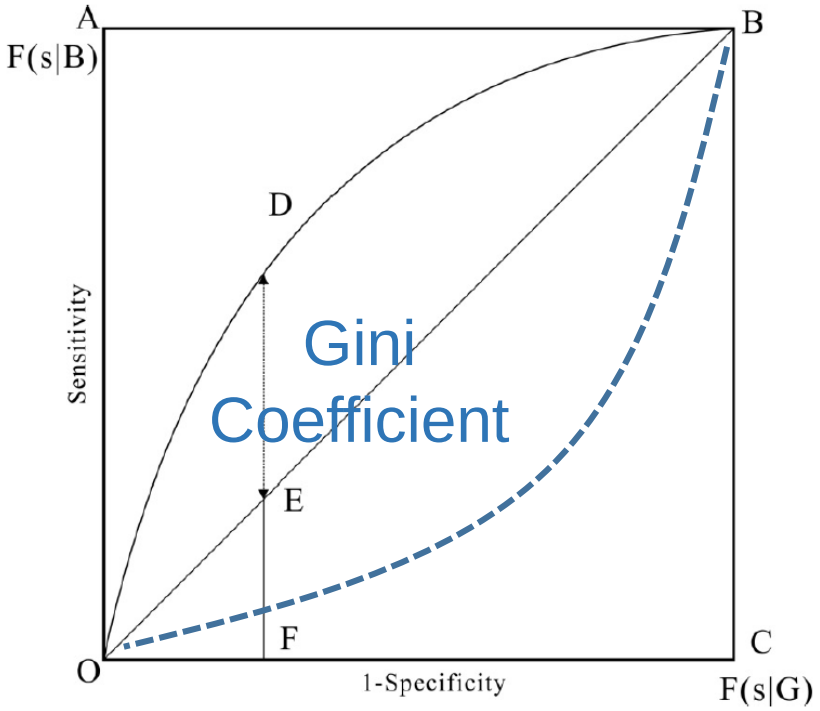
模型效果：Gini

- AUC的实际值分布在0.5~1之间，可以将AUC转换到0~1之间（与大部分指标分布一致），变成基尼系数。

- Gini系数：ROC曲线与对角线之间面积的两倍

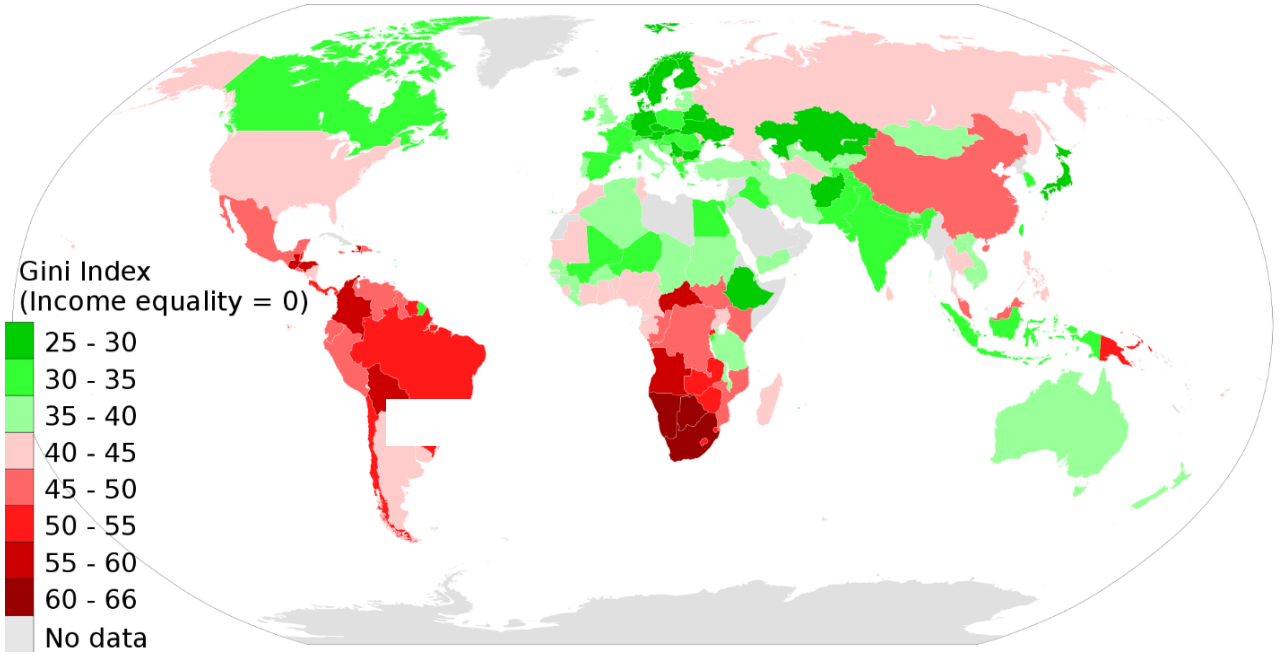
$$\text{Gini}=2\text{AUC}-1=2(\text{AUC}-0.5)$$

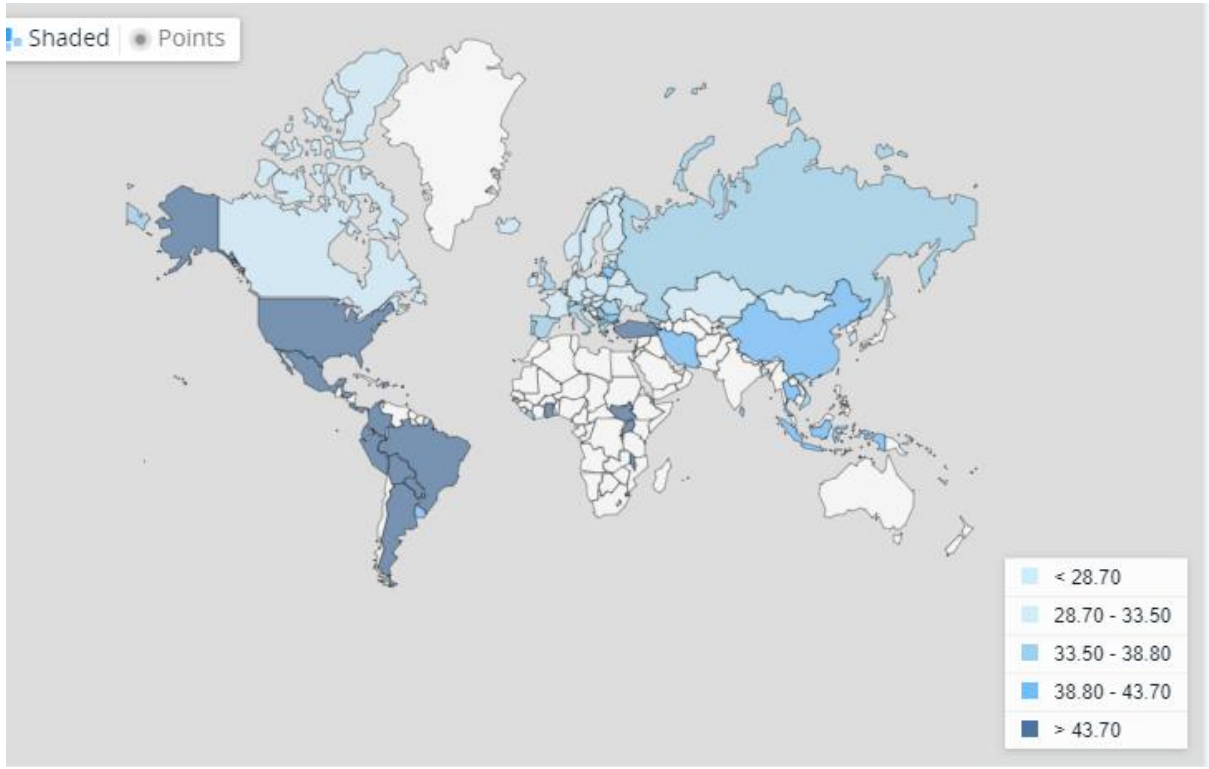
$$\begin{aligned}\text{Gini}&=2\int(F(s|\text{B})-F(s|\text{G}))dF(s|\text{G})\\&=2\int F(s|\text{B})dF(s|\text{G})-2\int F(s|\text{G})dF(s|\text{G})\\&=2\int F(s|\text{B})dF(s|\text{G})-\int d[F(s|\text{G})]^2\\&=2\text{AUC}-1\end{aligned}$$



模型效果：Gini

- 在国民收入中，基尼系数最大为“1”，最小为“0”。前者表示居民之间的年收入分配绝对不平均（即该年所有收入都集中在一个人手里，其余的国民没有收入），而后者则表示居民之间的该年收入分配绝对平均，即人与人之间收入绝对平等，基尼系数的实际数值只能介于这两种极端情况，即0～1之间。基尼系数越小，年收入分配越平均；基尼系数越大，年收入分配越不平均。
- 《中国家庭金融调查报告》：0.61@2010



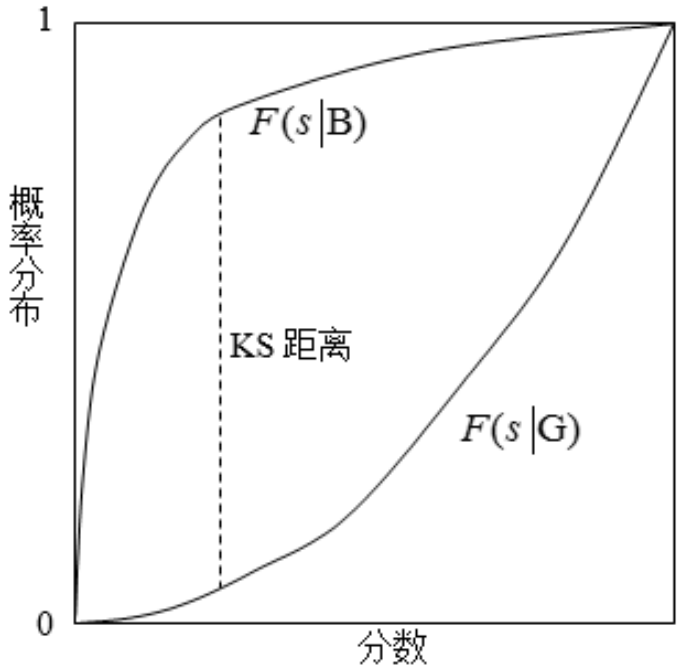


模型效果：KS

- KS统计量是分布函数 F(s|G), F(s|B) 间距离最大的垂直距离的长度。KS统计量又叫KS距离。
- $KS = \max_{S} |F(s|B) - F(s|G)|$

- KS取值范围0~1，距离越大，模型越好。

- KS统计量的缺点在于它描述的是在“最优分数”下的情形，但商业决策中需要一个相关或合适的决策分数。我们只能理解成，实际临界分数处的条件分布的距离比KS距离小，换句话说，KS统计量仅是好坏距离或区分度的上限。



模型效果：KS

- KS：Kolmogorov-Smirnov
- Kolmogorov 科尔莫哥洛夫（1903-1987）：几乎在数论之外的所有数学领域都做出杰出贡献。他热爱学生，对学生严格要求，指导有方，直接指导的学生有67人，他们大多数成为世界级的数学家，其中14人成为前苏联科学院院士。1963年，在第比利斯召开的概率统计会议上，美国统计学家沃尔夫维茨（1910-1981）说：“我来苏联的一个特别的目的是确定柯尔莫哥洛夫到底是一个人呢，还是一个研究机构。”
- KS在行业里最受喜欢，虽然99%的行业从业人员完全不理解KS距离的真实意义。
- “没有KS的信用评分是不完整的。甚至金融机构的高管们也熟悉它，并且为了相互攀比，通常号称自家模型的KS值比别家的高。”——Mays（2004）

模型效果：KS

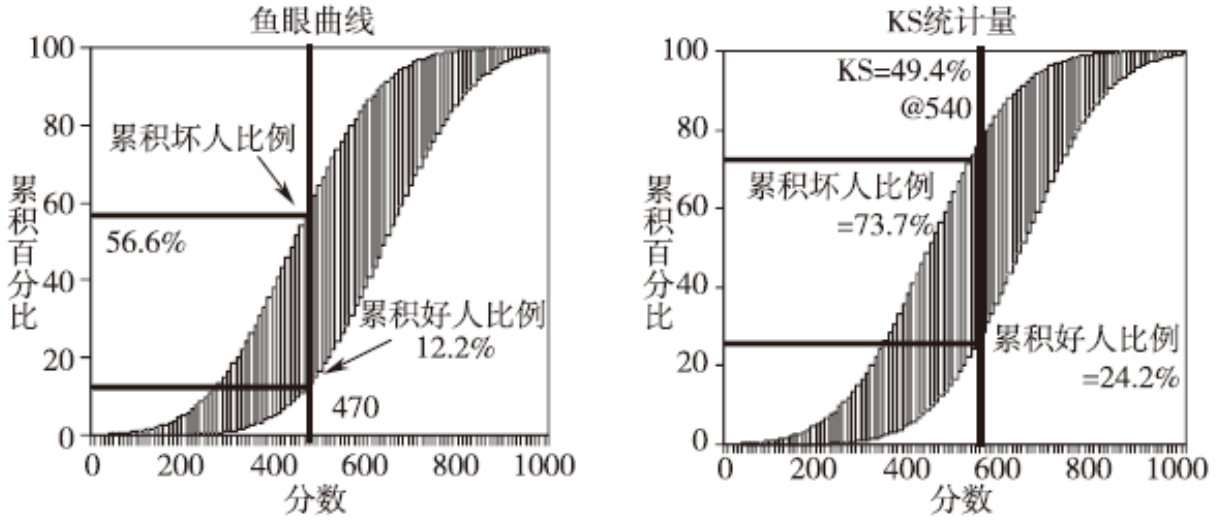


图 8.4　KS 曲线

KS 统计量就是两条曲线垂直距离绝对值最大的地方，所以有 $0<D_{KS}<1$。图 8.4 的右图中，470 分处的 KS 距离是 44.4%（56.6%−12.2%），在 550 分处增加到 49.4%。

式 8.5　KS 统计量　$D_{KS}=\max\{|cpY-cpX|\}$

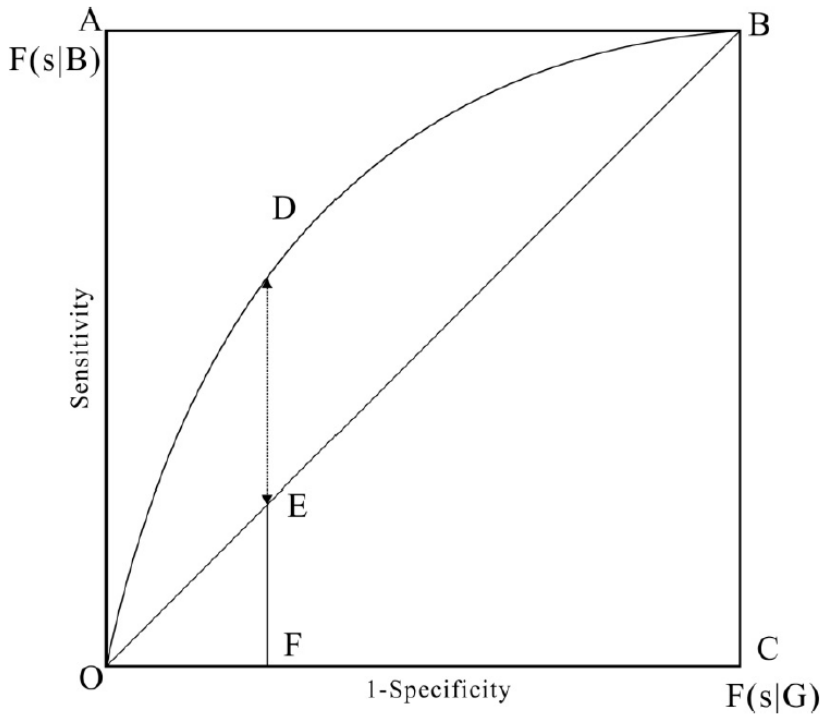
KS 统计量简单易懂，实际上可能又过于简单。KS 距离最大处对应曲线上的点可能和当前问题没有关联，尤其当其离当前临界线很远时，因此 KS 统计量通常和其他方法结合使用。

模型效果：KS

因为OF=EF

$$
\begin{aligned}
KS &= \max_s \left| F(s \mid B) - F(s \mid G) \right| \\
&= \max_s \left| DF - OF \right| \\
&= \max_s \left| DE + EF - OF \right| \\
&= \max_s \left| DE \right|
\end{aligned}
$$

KS距离变成了ROC曲线上的点到对角线的铅直距离最大的那个距离



模型效果：KS

- 计算KS

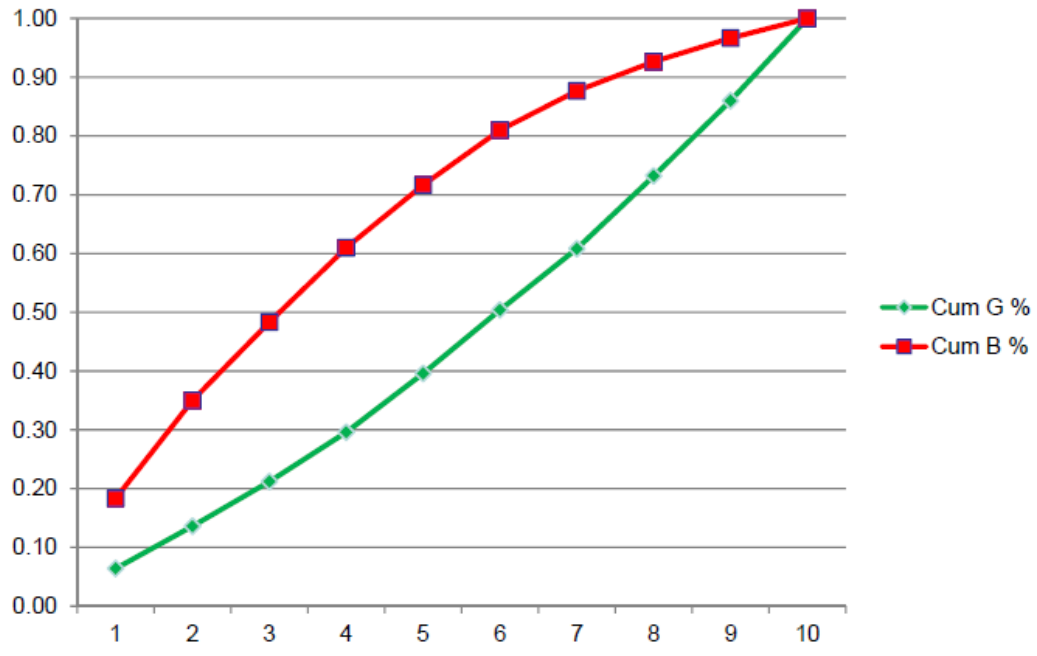
Score Band	< 0.5	0.5-0.55	0.56-0.6	0.61-0.65	0.66-0.7	0.71-0.75	0.76-0.8	0.81-0.85	0.86-0.9	> 0.9	Total
Goods	160	180	190	210	250	270	260	310	320	350	2500
Bads	55	50	40	38	32	28	20	15	12	10	300

模型效果：KS

- 计算KS

Score Band	Goods	Bads	Cum G Count	Cum B Count	Cum G Proportion $F_g(s)$	Cum B Proportion $F_b(s)$	$F_b(s) - F_g(s)$
< 0.5	160	55	160	55	0.06	0.18	0.12
0.5-0.55	180	50	340	105	0.14	0.35	0.21
0.56-0.6	190	40	530	145	0.21	0.48	0.27
0.61-0.65	210	38	740	183	0.30	0.61	0.31
0.66-0.7	250	32	990	215	0.40	0.72	0.32
0.71-0.75	270	28	1260	243	0.50	0.81	0.31
0.76-0.8	260	20	1520	263	0.61	0.88	0.27
0.81-0.85	310	15	1830	278	0.73	0.93	0.19
0.86-0.9	320	12	2150	290	0.86	0.97	0.11
> 0.9	350	10	2500	300	1.00	1.00	0.00

模型效果：KS



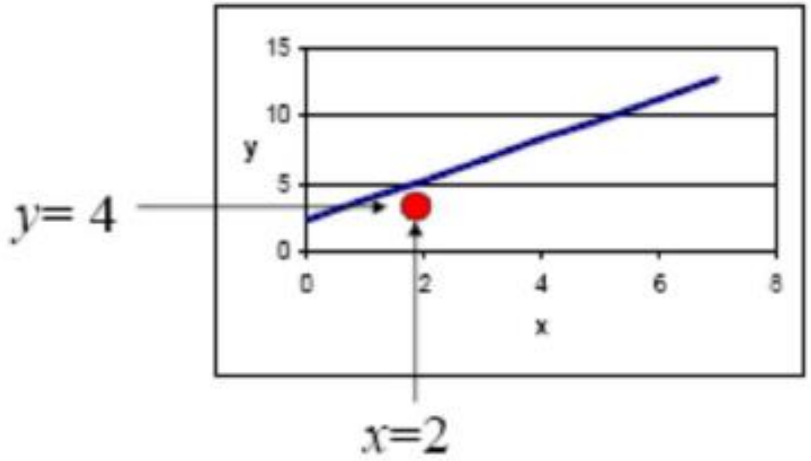
Score Band	Goods	Bads	Cum G	Cum B	Cum G %	Cum B %	K-S
below 0.5	0	300	0	300	0.00	1.00	1.00
0.5-0.55	0	0	0	300	0.00	1.00	1.00
0.56-0.6	0	0	0	300	0.00	1.00	1.00
0.61-0.65	0	0	0	300	0.00	1.00	1.00
0.66-0.7	0	0	0	300	0.00	1.00	1.00
0.71-0.75	0	0	0	300	0.00	1.00	1.00
0.76-0.8	0	0	0	300	0.00	1.00	1.00
0.81-0.85	0	0	0	300	0.00	1.00	1.00
0.86-0.9	0	0	0	300	0.00	1.00	1.00
above 0.9	2500	0	2500	300	1.00	1.00	0.00
Total	2500	300					1.00

- 理想的KS

线性回归

线性回归

- 一元一次回归： $y = \beta_0 + \beta_1 x + \varepsilon$
- 回归方程：$\hat{y} = b_0 + b_1 x$

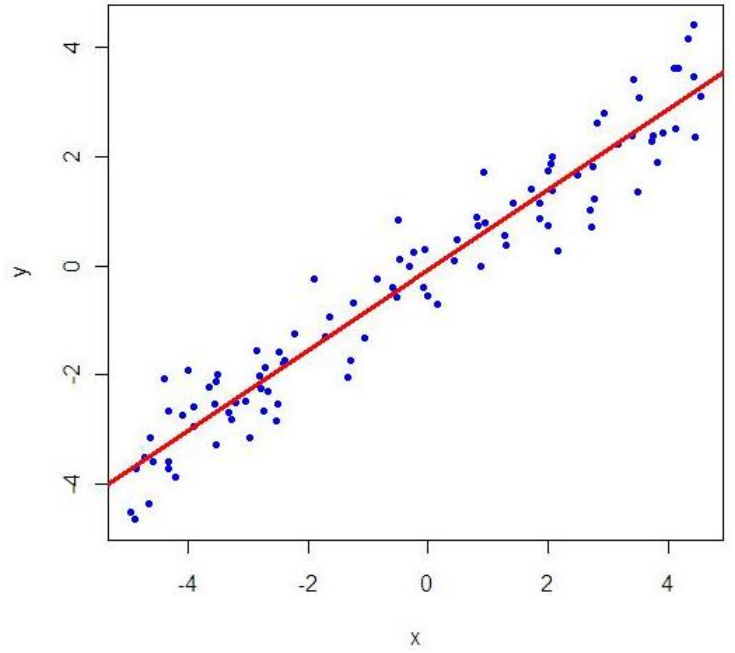


- 参数估计：OLS $\min \sum (y_i - \hat{y}_i)^2$
- 可决系数：

$$SST=SSR+SSE$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$r^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$



线性回归

多元线性回归

- 回归直线变成一个平板或更高维的图形

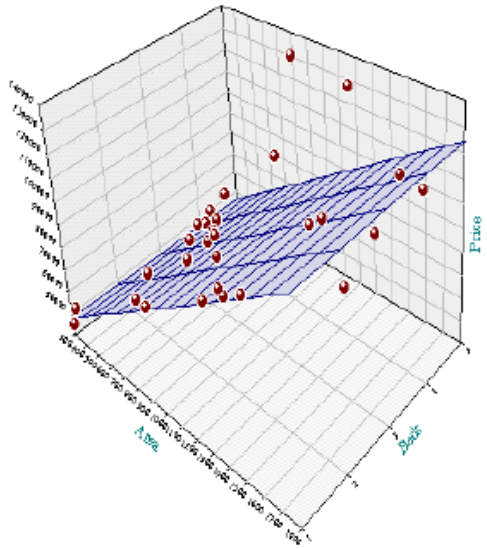
$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

- 检验r^2是否等于0：F 检验

$$F = \frac{SSR / k}{SSE / (n - k - 1)}$$

- 检验系数是否等于0：t 检验

$$H_0 : \beta_k = 0 \qquad t = \frac{b_k}{s_k}$$
$$H_1 : \beta_k \neq 0$$



线性回归

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.301[a]	.090	.052	.3914

a. Predictors: (Constant), loan_amnt, int_rate, annual_inc, fico score

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.447	4	.362	2.361	.059[b]
	Residual	14.553	95	.153		
	Total	16.000	99			

a. Dependent Variable: default

b. Predictors: (Constant), loan_amnt, int_rate, annual_inc, fico score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2.685	1.556		-1.726	.088
	fico score	.003	.002	.248	1.666	.099
	annual_inc	6.341E-8	.000	.004	.042	.966
	int_rate	.045	.015	.437	2.940	.004
	loan_amnt	-4.740E-6	.000	-.073	-.664	.508

a. Dependent Variable: default