

逻辑回归基础



分类变量之间的相关性检验




统计模型综述

预测变量的类型 反应变量的类型	分类	连续	连续和分类
连续	方差分析 (ANOVA)	普通最小二乘法 (OLS) 回归	协方差分析 (ANCOVA)
分类	列联表分析或 逻辑(Logistic) 回归	逻辑(Logistic) 回归	逻辑(Logistic) 回归

目标

- 创建双因素频数表
- 运行一个联合卡方检验
- 查看相关性的强度。
- 计算准确p值。
- 运行一个Mantel-Haenszel卡方检验。



分类变量间的相关性

- 如果一个分类变量的分布会随着另一个分类变量水平（值）的改变而改变，那么这两个分类变量间存在相关性。
- 如果不存相关性，那么无论另一个分类变量的水平（值）是多少，这一个变量的分布都相同。



无相关性



72%	28%
72%	28%

你们的情绪是否与天气情况相关呢？

存在相关性



18%	82%
40%	60%

你们的情绪是否与天气情况相关呢？

列联交叉表

- 列联交叉表显示了在每一行列组合上的观察值出现的频数。
被解释变量

		column 1	column 2	...	column c
解释变量	row 1	cell ₁₁	cell ₁₂	...	cell _{1c}
	row 2	cell ₂₁	cell ₂₂	...	cell _{2c}
	行百分比
	row r	cell _{r1}	cell _{r2}	...	cell _{rc}

结果展示

话费上升、流失表			
话费上升	是否流失		
行（百分比）	否	是	总计
否	45.57%	54.43%	N=1819
是	66.91%	33.09%	N=1644
总	N=1929	N=1534	N=3463

对于二乘二的分类变量, 只要看一列就可以。上面的结果表明话费上涨的客户, 不流失的比例更大。

相关性检验

原假设

- 话费上升和是否流失之间不存在相关性。
- 无论统计期间话费是否上升，其流失的可能性是相同的。

• 备择假设

- 话费上升和是否流失之间存在相关性。
- 统计期间话费是否上升影响流失的可能性。

卡方检验

不相关

观察频数 = 期望频数

相关

观察频数 \neq 期望频数



期望频数由此式子计算出：



$(\text{row total} * \text{column total}) / \text{sample size}.$

(行总计 * 列总计) / 样本量

也可以，单元格的期望频数 = 行百分比 * 列百分比 * 样本总数

其中：行百分比也被称为行边际分布

卡方检验

卡方检验和对应的P值

- 确定相关性是否存在
- 不能测量相关性的强弱
- 取决于样本量，并反映样本量。

$$\chi^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}}$$

优势比（Odds Ratios）

- 优势比反映了从发生比来看，一个特定事件在一个组发生的可能性相对于在另一个组发生的可能性的的大小？

$$\text{Odds} = \frac{p_{event}}{1 - p_{event}}$$

结果的概率 vs 结果的发生比

	结果		总计
	否	是	
组 A	20	60	80
组 B	10	90	100
总计	30	150	180

B组中结果为“是”
的总计

÷

B组结果总计

B组中“是”的概率 = 90 ÷ 100 = 0.9

结果的概率 vs 结果的发生比

	结果		总计
	否	是	
组 A	20	60	80
组 B	10	90	100
总计	30	150	180

B组中“是”的概率=
0.90

÷

B组中“否”的概率=
0.10

B组中“是”的发生比= **0.90 ÷ 0.10 = 9**

Odds Ratio

	结果		总计
	否	是	
组 A	20	60	80
组 B	10	90	100
总计	30	150	180

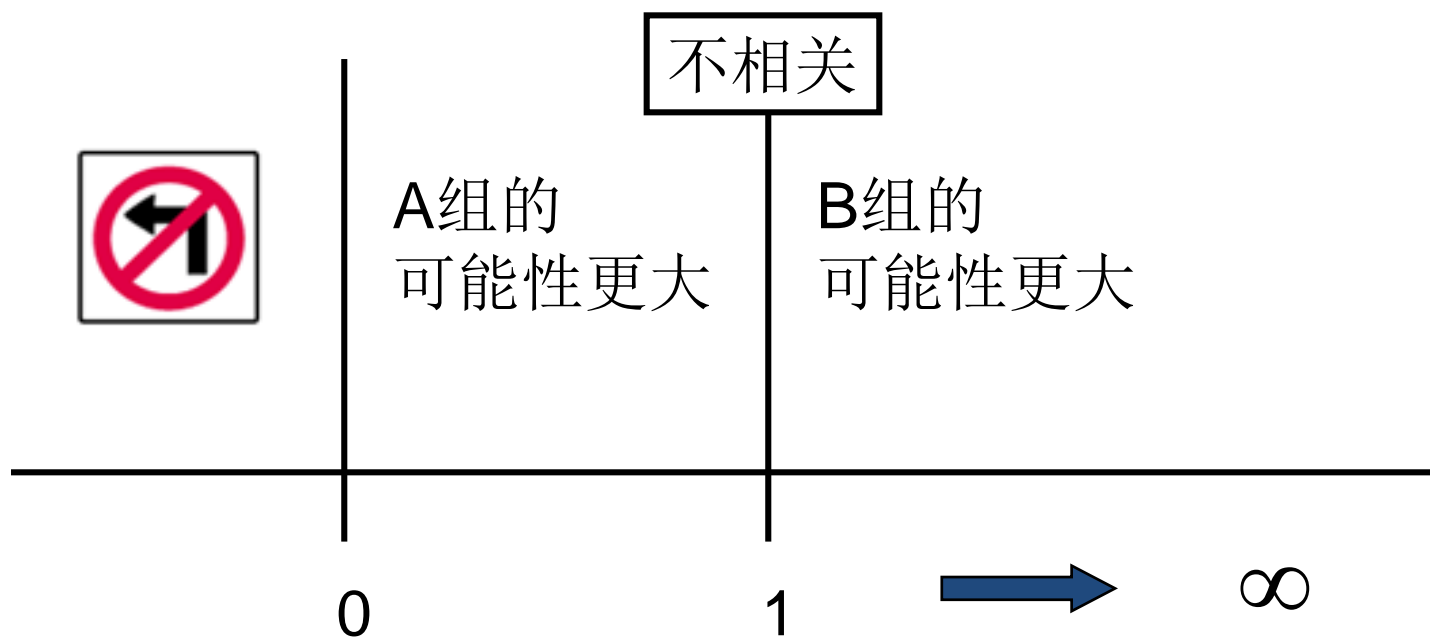
B组中“是”
的发生比= 9

÷

A组中“是”的
发生比= 3

Odds Ratio, B 对 A = 9 ÷ 3 = 3

Odds Ratio (B 对 A) 的属性



优势比的实际运用

德国赢的概率： $60/(60+20)=0.75$

德国输的概率： $1-0.75=0.25$

德国的Odds= $0.75/0.25=3$

巴西赢的概率： $90/(90+10)=0.9$

巴西输的概率： $1-0.9=0.1$

巴西的Odds= $0.9/0.1=9$

巴西对德国的Odds Ratios= $3/1=3$

	赢	输
德国	60	20
巴西	90	10

西方赌球玩法：如上，经过计算得到巴西对德国的胜率为0.75，故巴西队赔率为 $1:0.75=1.3$ ；同理可得德国队赔率为4，还存在踢平的情况，这里省略，则：

如果下注1元赌巴西赢，则结果巴西赢了可得1.3元，输赔本。

如果下注1元赌德国赢，则结果德国赢了可得4元，输赔本。

逻辑回归入门

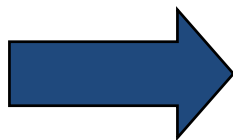
目标

- 定义逻辑回归的概念。
- 使用逻辑回归任务拟合一个二元逻辑回归模型。
- 描述一个含有连续预测变量的逻辑回归任务的标准输出。
- 阅读和解释odds ratio 表和图。

客户流失的预测

预测变量

在网时长



反应变量

客户离网



统计模型综述

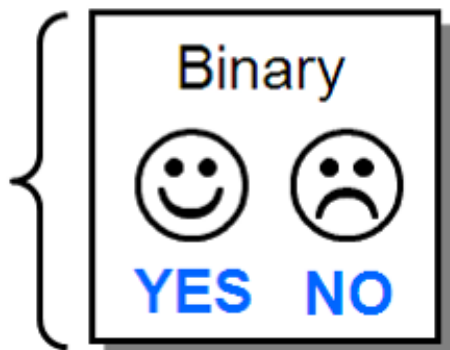
预测变量的类型 反应变量的类型	分类	连续	连续和分类
连续	方差分析 (ANOVA)	普通最小二乘法 (OLS) 回归	协方差分析 (ANCOVA)
分类	列联表分析或 逻辑(Logistic) 回归	逻辑(Logistic) 回归	逻辑(Logistic) 回归

逻辑回归的类型

反应变量

逻辑回归的类型

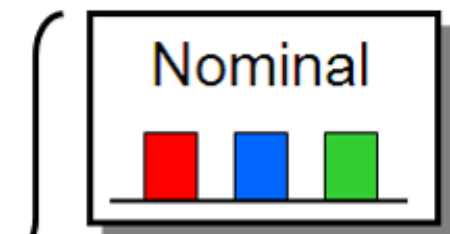
两类



二元

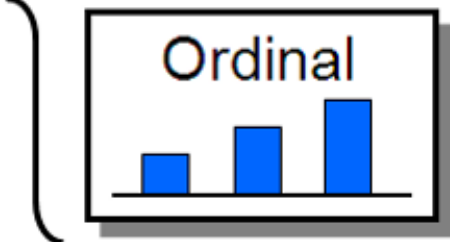
Binary

三类或更多



名义

Nominal



序数

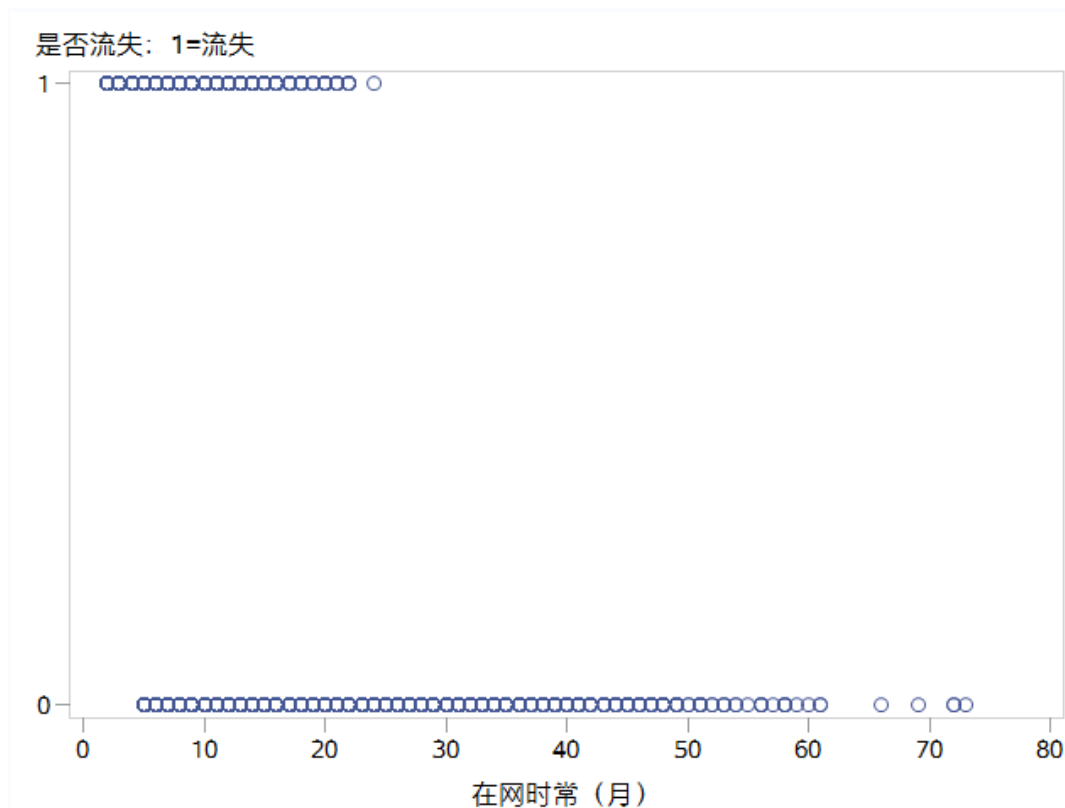
Ordinal

什么不用普通最小二乘法回归？

$$OLS \text{ 回归: } Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

- 如果反应变量是分类变量，如何将其编码为数值型？
- 如果反应变量被编码为：1=是、0=否，当你的回归式预测结果是0.5、1.1或-0.4时，结果的实际意义是什么？
- 如果只有两个（或一些）可能的反应水平，假设方差不变以及正态分布是合理的吗？

在网时长vs流失与否的散点图



线性概率模型？

线性概率模型: $p_i = \beta_0 + \beta_1 X_{1i}$

- 概率是有界的，但线性函数可取任何值。（当预测值为 -0.4 或 1.1 时，你将如何解释？）
- 给定概率的有界性，你能否通过限定 X 的可能范围来假设一个 X 和 p 的线性关系？
- 能否假设一个方差不变的随机误差？
- 一个观察值的观测概率是什么？

逻辑回归模型与Logit 转换

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki}$$

$\text{logit}(p_i)$ = 事件发生概率的logit

β_0 = 回归式的截距

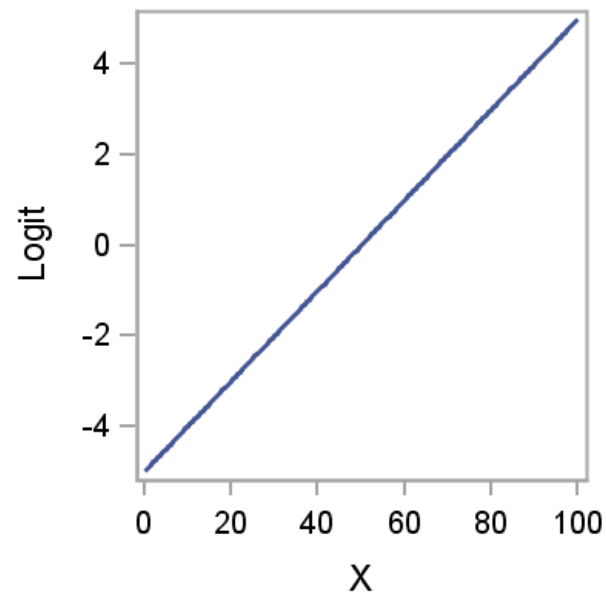
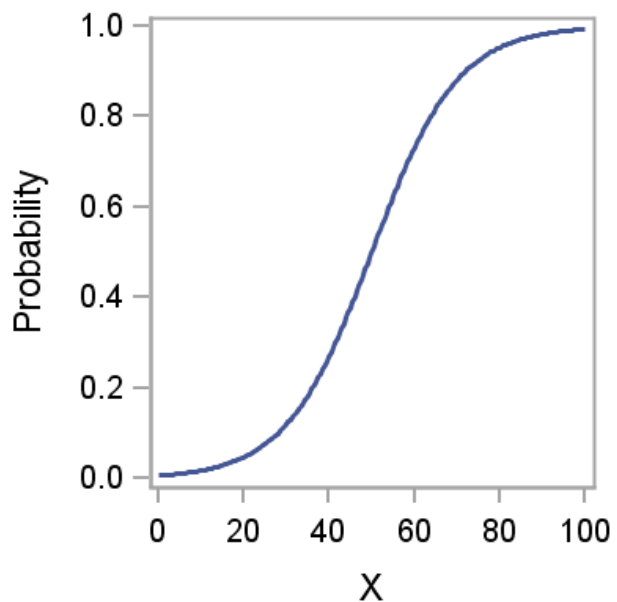
β_k = 第k个预测变量的参数估计

$$\text{logit}(p_i) = \ln \left(\frac{p_i}{(1 - p_i)} \right)$$

- *Logit*是发生比 (*odds*) 的自然对数
- i 表示所有案例(观察值)
- p_i 在第*i*个案例中一个事件发生的概率
- \ln 是自然对数 (底数为e)

示意

Logit转换



$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

上面假设的另一种理解方式

大家考虑过买IPAD吗？

假设Ipad的价格是2000元，而每个人内心对他的评价是 \tilde{Y}

$$\left\{ \begin{array}{ll} Y = 1 & \text{如果 } \tilde{Y} \geq 2000 \\ Y = 0 & \text{如果 } \tilde{Y} < 2000 \end{array} \right.$$

因此，可以 $\tilde{Y}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ ，而 \tilde{Y} 是Y的一种转换形式。

\tilde{Y} 是连续的，取值范围为 $(-\infty, +\infty)$ ，除此之外，没有任何限制，可以是任意的函数，服从任意分布。

- 第一个想到： $P(Y=1)$ ，值域是 $[0, 1]$
- 其次是： $P(Y=1) / [1 - P(Y=1)]$ ，值域是 $[0, +\infty]$
- 最后，取个自然对数： $\ln\{P(Y=1) / [1 - P(Y=1)]\}$ ，值域是 $[-\infty, +\infty]$

结果说明

模型总体
显著度

检验全局零假设: BETA=0			
检验	卡方	自由度	Pr > 卡方
似然比	1723.2836	1	<.0001
评分	1114.8382	1	<.0001
Wald	698.2095	1	<.0001

逻辑回归没有提供R方，因此没有办法知道解释变量解释了变异的百分比。提供了三种极大似然估计常用的统计量，当三个都显著时，说明至少有一个解释变量具有解释力度。

变量每个
水平的估计
系数

最大似然估计分析					
参数	自由度	估计	标准 误差	Wald 卡方	Pr > 卡方
Intercept	1	2.5882	0.1004	663.9175	<.0001
duration	1	-0.2482	0.00939	698.2095	<.0001

这里的“估计”等同于回归结果，负的回归系数表明在网时长越长，流失的可能性越小。但是具体的数值意义不容易理解，虽然下面会讲解，但是关注这个数值的意义不大。

每个变量一
单位变化的
优势比

优比估计			
效应	点估计	95% Wald 置信限	
duration	0.780	0.766	0.795

优势比大于1，表明随着该解释变量的提高，Y=1的概率在增大。

预测值

根据右侧的公式给每个观测预测 $y=1$ 的概率

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

⑫ 个人客户的ID	⑫ 单个概率: churn=1
74961005	0.9994583966
74929842	0.9994397695
74808121	0.9991592587
74919003	0.9984661971
74855032	0.9984524769
74842778	0.9983242248
74841623	0.9981002856
74861472	0.9980441824
73417274	0.9979985216
74158095	0.9979708375
74115753	0.9974554829
74816707	0.9973049003
74927592	0.996768544
74966972	0.9963245117
74236709	0.9963224259
74932993	0.9961538611



逻辑回归模型

- 该示例阐述了之前讨论的概念。

从现有逻辑回归模型中计算优势比

- 逻辑回归模型:

$$\text{logit}(\hat{p}) = \ln(\text{odds}) = \beta_0 + \beta_1 \times (\text{duration})$$

- Odds ratio (在网时常相差1月):

$$\text{odds}_{\text{长}} = e^{\beta_0 + \beta_1 \times (\text{duration} + 1)}$$

$$\text{odds}_{\text{短}} = e^{\beta_0 + \beta_1 \times (\text{duration})}$$

$$\text{Odds Ratio} = \frac{e^{\beta_0 + \beta_1 \times (\text{duration} + 1)}}{e^{\beta_0 + \beta_1 \times (\text{duration})}} = e^{\beta_1}$$

$$= e^{(-0.248)} = 0.78$$



估计方法介绍

极大似然估计：我们已经收集到由 (x_i, y_i) 组成的样本，我们不知道的是描述 x 和 y 之间关系的参数，和 y 分布的参数。虽然我们不知道这些参数，但是我们可以对这些参数的性质进行假设。1）这些参数应该使得我们得到 (x_i, y_i) 这组样本的可能性最大；2） y 要符合某一个给定分布，比如扰动项服从正态分布（线性回归）、扰动项服从逻辑分布（逻辑回归）。

极大斯然估计的推导思路

大家考虑过买IPAD吗？

假设Ipad的价格是2000元，而每个人内心对他的评价是 \tilde{Y}

$$\begin{cases} Y = 1 & \text{如果 } \tilde{Y} \geq 2000 \\ Y = 0 & \text{如果 } \tilde{Y} < 2000 \end{cases}$$

因此，可以 $\tilde{Y}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ ，而 \tilde{Y} 是Y的一种转换形式。

\tilde{Y} 是连续的，取值范围为 $(-\infty, +\infty)$ ，除此之外，没有任何限制，可以是任意的函数，服从任意分布。

- 第一个想到： $P(Y=1)$ ，值域是 $[0, 1]$
- 其次是： $P(Y=1) / [1 - P(Y=1)]$ ，值域是 $[0, +\infty]$
- 最后，取个自然对数： $\ln\{P(Y=1) / [1 - P(Y=1)]\}$ ，值域是 $[-\infty, +\infty]$

逻辑回归的方法推导：极大似然估计

假设我们在推销Ipad，每个消费者都有一个效用函数，消费者对Ipad的需求受一些解释变量的影响，比如阅读的次数、玩游戏的次数等等。我们 y^* 用来代表效用函数，它是 \mathbf{x} 的线性函数。

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon.$$

但是我们作为商家，是不知道消费者的效用函数的，我们只能知道他是否购买Ipad，用 y 来表示观测结果。为了简单起见，我们假设ipad的价格为0.

$$y = \begin{cases} 1 & \text{if } y^* > 0, \\ 0 & \text{if } y^* \leq 0. \end{cases}$$

逻辑回归的方法推导：极大似然估计

这里面 y^* 被称为隐变量（latent variable）。接下来我们构造似然函数。
购买Ipad客户的概率为：

$$\begin{aligned}\Pr(y = 1|\beta, \sigma^2, \mathbf{x}) &= \Pr(y^* > 0|\beta, \sigma^2, \mathbf{x}) \\ &= \Pr(\mathbf{x}'\beta + \varepsilon > 0|\beta, \sigma^2, \mathbf{x}) \\ &= \Pr(\varepsilon > -\mathbf{x}'\beta|\beta, \sigma^2, \mathbf{x}) \\ &= 1 - F(-\mathbf{x}'\beta),\end{aligned}$$

其中 $F(\cdot)$ 为扰动项 ε 的累积概率密度函数。
不购买Ipad客户的概率为：

$$\begin{aligned}\Pr(y = 0|\beta, \sigma^2, \mathbf{x}) &= 1 - \Pr(y = 1|\beta, \sigma^2, \mathbf{x}) \\ &= F(-\mathbf{x}'\beta).\end{aligned}$$

逻辑回归的方法推导：极大似然估计

得到似然函数为：

$$\prod_{y=0} F(-\mathbf{x}'\boldsymbol{\beta}) \prod_{y=1} [1 - F(-\mathbf{x}'\boldsymbol{\beta})] \quad \text{——} \quad (1\text{式})$$

当假设扰动项 $\boldsymbol{\varepsilon}$ 服从逻辑分布时，则累积概率密度函数为：

$$F(-\mathbf{x}'\boldsymbol{\beta}) = \frac{\exp(-\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(-\mathbf{x}'\boldsymbol{\beta})} = \frac{1}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})} \quad \text{——} \quad (2\text{式})$$

和

$$1 - F(-\mathbf{x}'\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})} \quad \text{——} \quad (3\text{式})$$

将（2式）和（3式）带入（1式），则得到逻辑回归的似然函数。其求解过程和线性回归的极大似然估计完全一样。

逻辑回归的方法推导：极大似然估计

得到对数似然函数为：

$$\ln L(Y, \beta) = \sum_{i=1}^n y_i x_i^T \beta - \sum_{i=1}^n \ln [1 + \exp(x_i^T \beta)]$$

对参数求偏导为：

$$\frac{d \ln L(Y, \beta)}{d \beta} = \sum_i y_i x_i^T - \sum_i \left[\frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \right] x_i^T = 0$$

不过这个上面的这个式子没有解析解，一般使用Newton-Raphson方法进行数值计算。

正则化的逻辑回归


岭回归（L2正则）：

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^n \log(\exp(-y_i(X_i^T w + c)) + 1)$$

Lasso（L1正则）：

$$\min_{w,c} \|w\|_1 + C \sum_{i=1}^n \log(\exp(-y_i(X_i^T w + c)) + 1)$$

模型表现优劣的评估



模型评估：成对比较

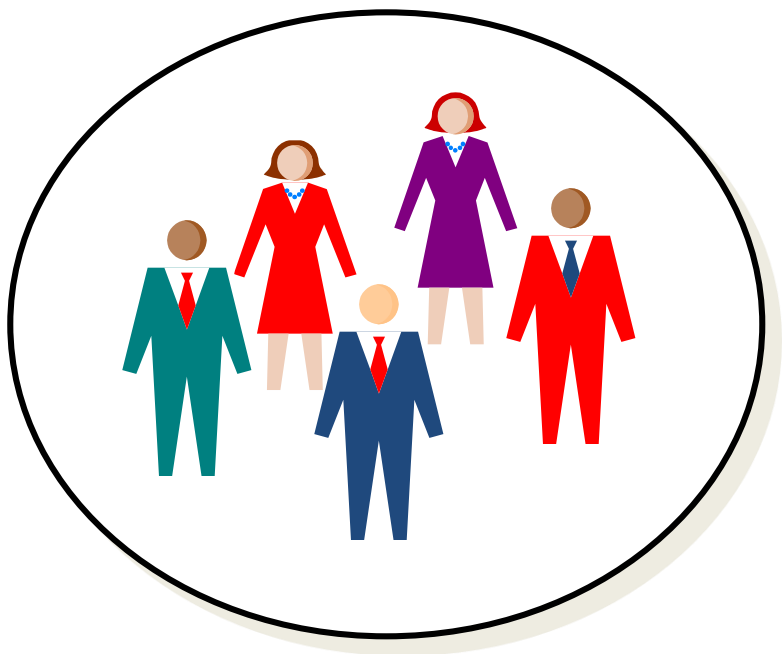
- 计算一致的对数、不一致的对数以及相等（tied）的对数来评估模型是否很好地预测了自身的数据，从而判断模型拟合得优秀与否。
- 通常我们希望一致对的占比高，不一致对和相等对的占比低。



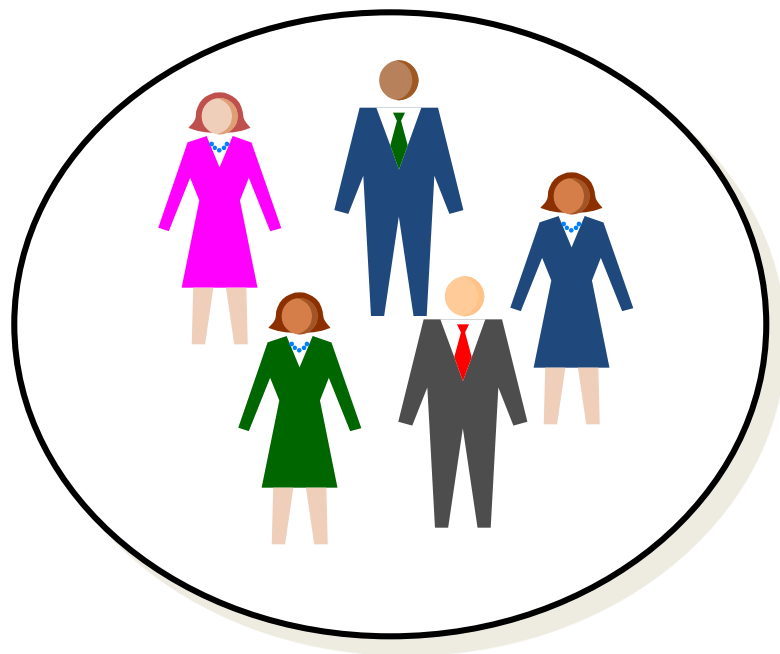
成对比较

为了找到一致对、不一致对以及相等对，要将每一个获得相关结果的人与每一个没有获得相关结果的人进行比较。

流失



保留



一致对

在网时长和客户保留正相关，因此在网时长越长的， P （保留）的预测值越大。

比较一个在网时长6个月的保留客户和一个在网时长3个月的流失客户。

流失， 3个月



$$P(\text{保留}) = .4677$$

保留， 6个月



$$P(\text{保留}) = .5272$$

实际的排序与模型相符。这是一个一致对。

不一致对

比较一个在网时长2个月的保留客户和一个在网时长3个月的流失客户。

流失， 3个月



$P(\text{保留}) = .4677$

保留， 2个月



$P(\text{保留}) = .4091$

实际排序与模型不符。这是一个不一致对。

相等对

比较两个在网时长1个月，一个流失，一个保留。

保留， 1个月



$$P(\text{保留}) = .3697$$

流失， 1个月



$$P(\text{保留}) = .3697$$

模型不能分辨其二者。这是一个相等对。

模型：一致对、不一致对和相等对

预测概率和观测响应的关联			
一致部分所占百分比	86.1	Somers D	0.747
不一致部分所占百分比	11.3	Gamma	0.767
结值百分比	2.6	Tau-a	0.369
对	2959086	c	0.874



Question & Answer

8.3 模型评估

样本内评估

样本内评估：使用训练集同期的数据



样本外评估：使用下一期的滚动数据



评估指标汇总

预测类型	统计量
决策 (Decisions)	准确率/误分类 利润/成本
排序 (Rankings)	ROC 指标 (一致性) Gini 指数 K-S统计量 提升度

决策类模型评估

该类模型的需求是回答“是不是？”。比如判别持身份证办业务的人是否为证件所有者。

混淆矩阵： 每给定一个阈值，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

1. 正确率 = $(A+D)/(A+B+C+D)$
2. 灵敏度（**Sensitivity**；覆盖率recall）= $A/(A+B)$
3. 命中率(Precision、PV+)= $A/(A+C)$
- 特异度 (**Specificity**；负例的覆盖率)= $D/(C+D)$
5. 负命中率(PV-) = $D/(D+B)$

评估指标汇总

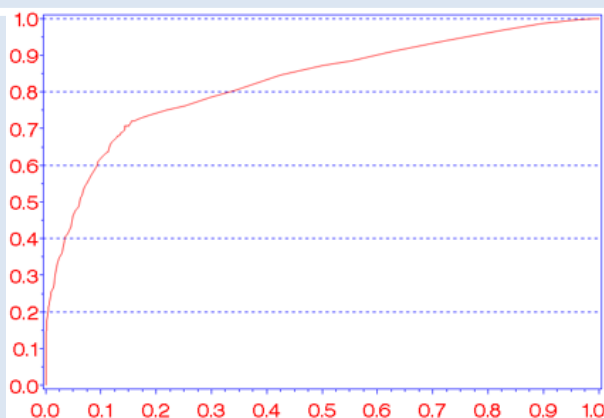
预测类型	统计量
决策 (Decisions)	精确性/误分类 利润/成本
排序 (Rankings)	ROC 指标 (一致性) Gini 指数 K-S统计量 提升度

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

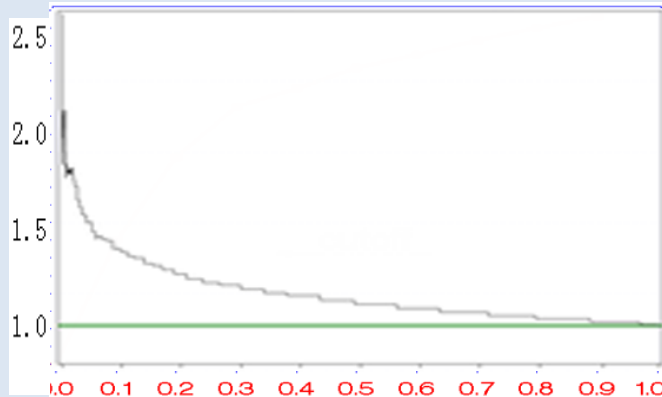
ROC曲线：用来描述模型分辨能力,对角线以上的图形越高模型越好

X:1-特异度
Y: 灵敏度



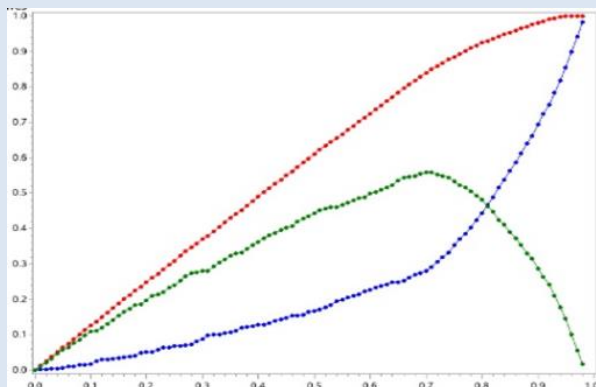
累积提升曲线：由于展示使用模型预测结果与随机情况下获取显性样本的能力比较

X:深度
Y: 正例的
累积密度
除以基准
概率



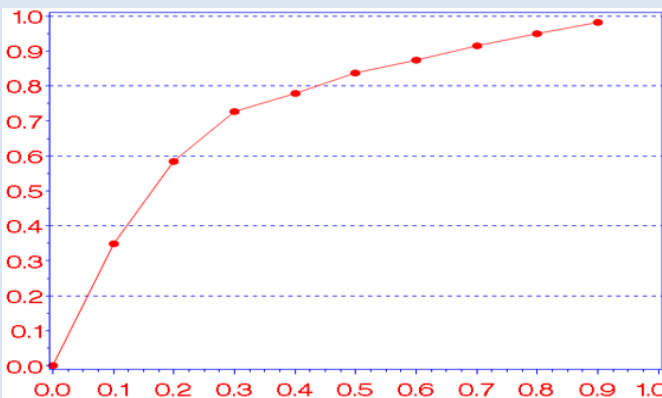
K-S曲线：用来描述模型对违约客户的分辨能力

X:深度
Y红：正例的
累积密度
Y蓝：负例的
累积密度
Y率：K-S值



洛伦兹曲线：用来描述预期违约客户的分布

X:深度
Y: 正例的
累积密度



模型评估——ROC曲线

ROC (Receiver Operating Characterstic) 曲线——接收者操作特征曲线。

最早应用于雷达信号检测领域，用于区分信号与噪声。

信号检测论：在听觉感受性相同的情况下，判断标准不一样。

①冒进：每次出现不会“漏报”，感觉有就报告。

②保守：每次出现不会“虚报”，没有把握不会报告

。

模型评估——ROC曲线

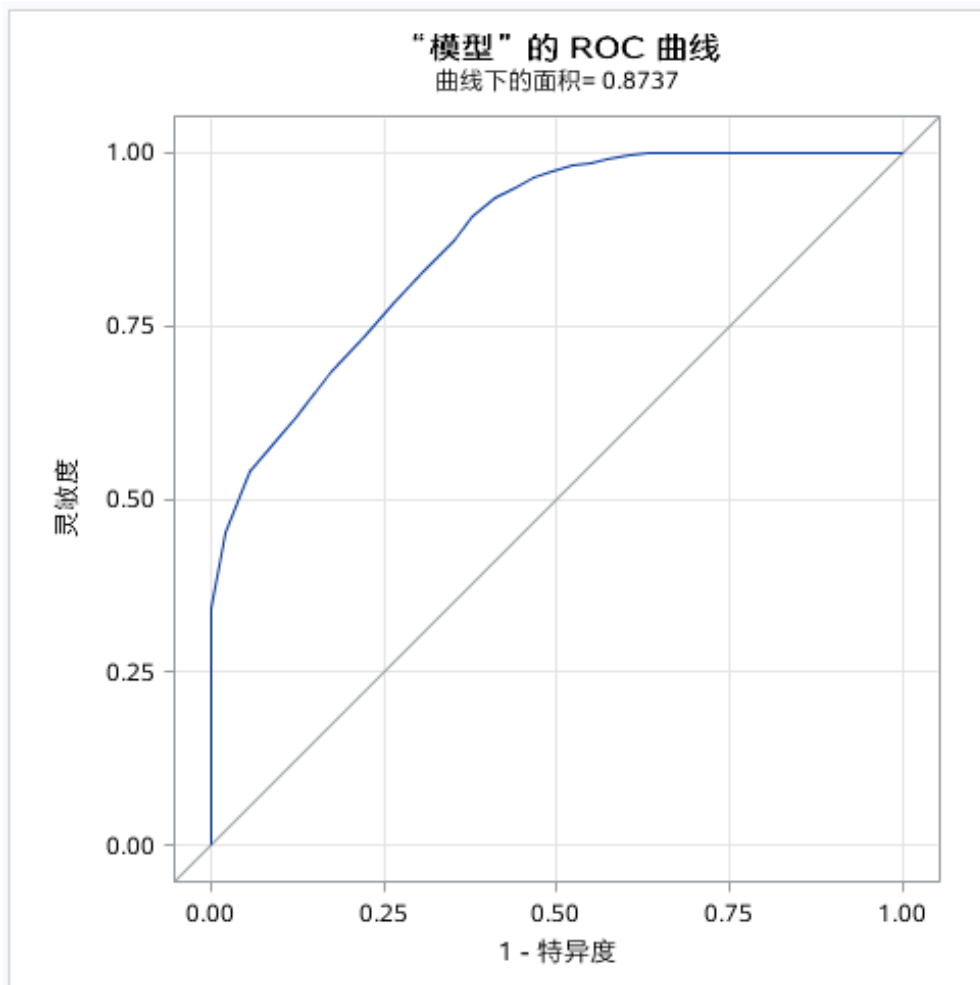
混淆矩阵：每给定一个阈值，就可以做出一个混淆矩阵

		打分值		合计
		反应（预测=1）	未反应（预测=0）	
真实结果	呈现信号 （真实=1）	A（击中）	B（漏报）	A + B
	未呈现信号 （真实=0）	C（虚报）	D（正确否定）	C + D
合计		A + C	B + D	A + B + C + D

灵敏度= $A / (A + B)$

特异度 = $D / (C + D)$

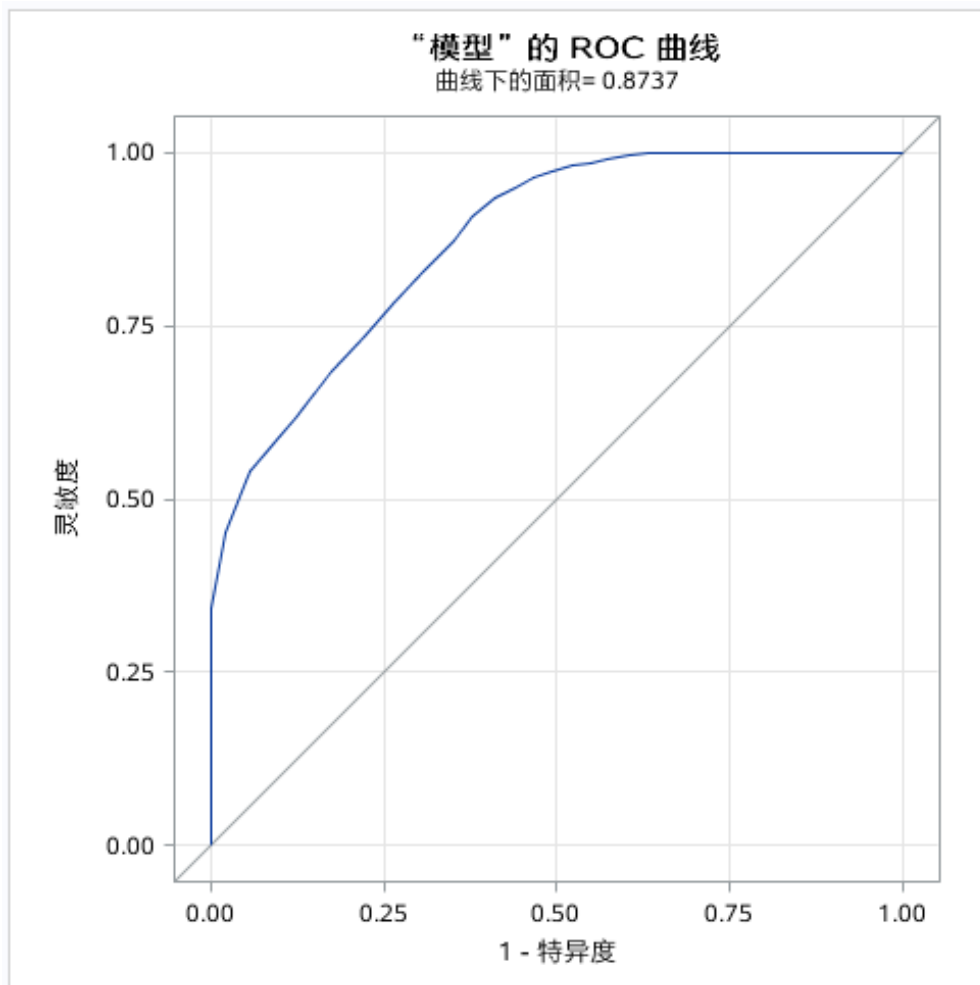
模型评估——ROC曲线



特征:

- ①灵敏度和特异性越大越好。
- ②ROC曲线间的相对重要性。
- ③沿对角线分布可以认为是随机因素造成的。

ROC曲线



特征:

- ①敏感度和特异性越大越好
- ②ROC曲线间的相对重要性
- ③沿对角线分布可以认为是随机因素造成的。

ROC图形制作方法1

混淆矩阵：

CM		真实	
		0	1
预测	0	8	8
	1	1	3

Accuracy = 0.55
Precision = 0.75
Recall = 0.27

Sensitivity = Recall

Specificity
= True Negative Rate
= 0.89

Predicted Probability	True Class	Sensitivity	Specificity	1-Specificity
0.90	1	0.09	3-1.00	0.00
0.80	1	0.18	3-1.00	0.00
0.70	0	0.18	0.89	0.11
0.60	1	0.27	0.89	0.11
0.55	1	0.36	0.89	0.11
0.54	1	0.45	0.89	0.11
0.53	1	0.55	0.89	0.11
0.52	0	0.55	0.78	0.22
0.51	1	0.64	0.78	0.22
0.51	1	0.73	0.78	0.22
0.40	1	0.82	0.78	0.22
0.39	0	0.82	0.67	0.33
0.38	1	0.91	0.67	0.33
0.37	0	0.91	0.56	0.44
0.36	0	0.91	0.44	0.56
0.35	0	0.91	0.33	0.67
0.34	1	3-1.00	0.33	0.67
0.33	0	3-1.00	0.22	0.78
0.30	0	3-1.00	0.11	0.89
0.10	0	3-1.00	0.00	3-1.00

0.555

0

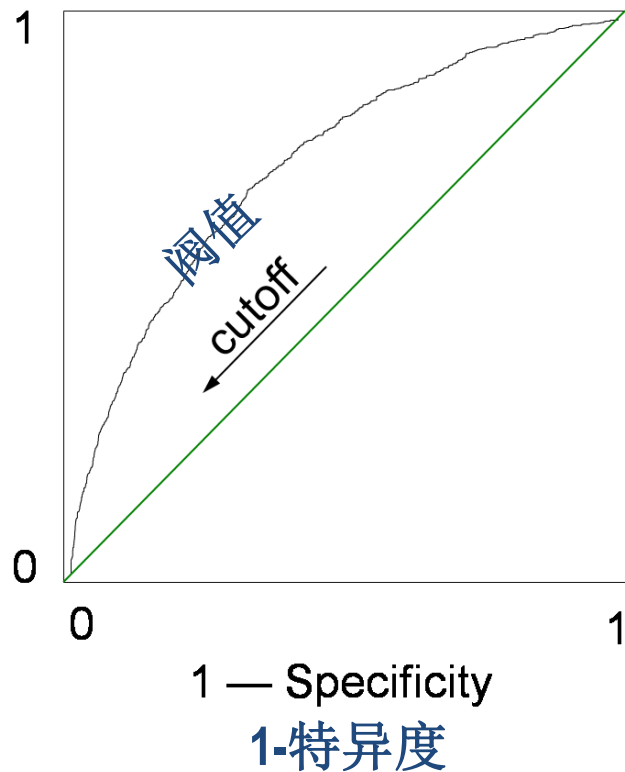
1

ROC图形制作方法2

	⑫ 阈值	⑫ 敏感度	⑫ 特异度
1	0.96	0.003	0.998
2	0.91	0.022	0.997
3	0.89	0.038	0.996
4	0.83	0.082	0.989
5	0.75	0.173	0.959
6	0.72	0.227	0.948
7	0.69	0.280	0.929
8	0.61	0.438	0.846
9	0.60	0.467	0.834
10	0.52	0.670	0.721
11	0.48	0.730	0.666
12	0.41	0.829	0.517
13	0.36	0.880	0.412
14	0.35	0.882	0.399
15	0.30	0.908	0.323
16	0.21	0.953	0.200
17	0.17	0.967	0.152
18	0.11	0.983	0.092
19	0.05	0.991	0.048
20	0.00	0.998	0.005

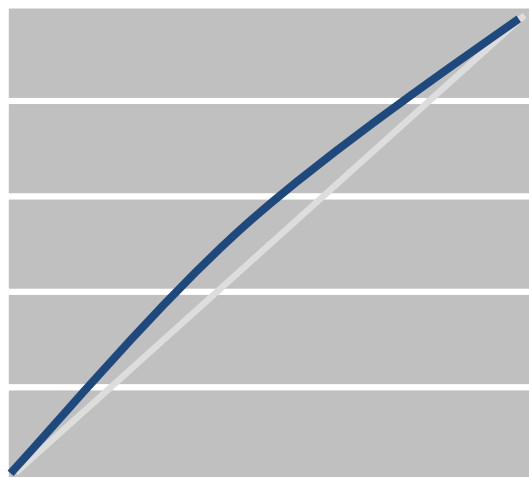
阈值下降

灵敏度
Sensitivity

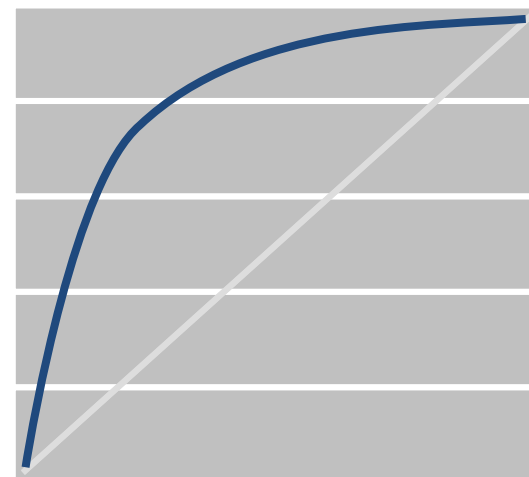
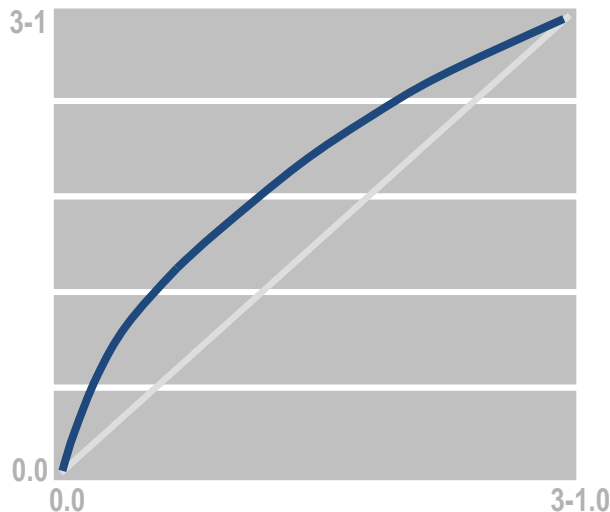


随着阈值的下降，灵敏度在升高，特异度在降低。

ROC 图形使用



弱的模型



强的模型

ROC曲线结果的取值在 $[0.5, 1]$ 。

一般来说，

$[0.5, 0.7)$ 表示效果较低, 但是预测股票已经很不错了;

$[0.7, 0.85)$ 表示效果一般;

$[0.85, 0.95)$ 表示效果良好;

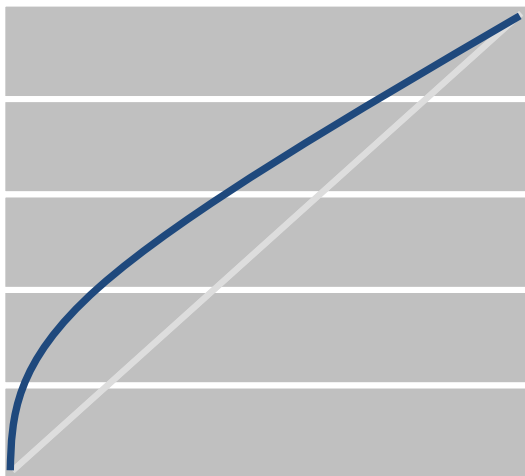
$[0.95, 1]$ 社会科学建模中不大可能出现。

注意:

①有时ROC曲线可能会落入对角线以下, 这时需检查检验方向与状态值的对应关系

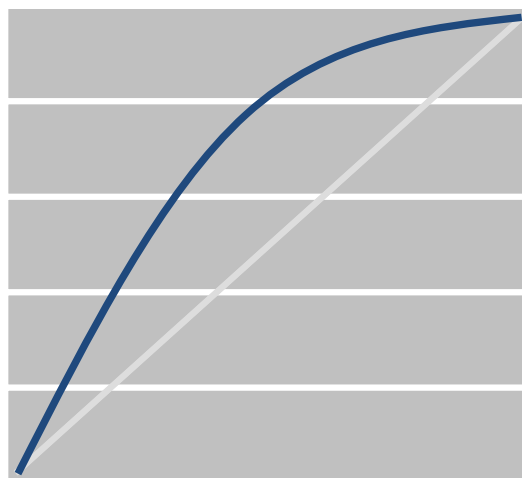
②如果某ROC曲线在对角线两边均有分布, 需检查数据或专业背景。

ROC 图形



违约分值高处敏感

该模型在违约风险**高**人群中的预测能力较强，而在违约率**低**的部分较弱。有些业务需要做出这样的模型，比如汽车金融公司，业务需要只把违约风险非常高的客户筛选出来，而大部分客户授予分期贷款。

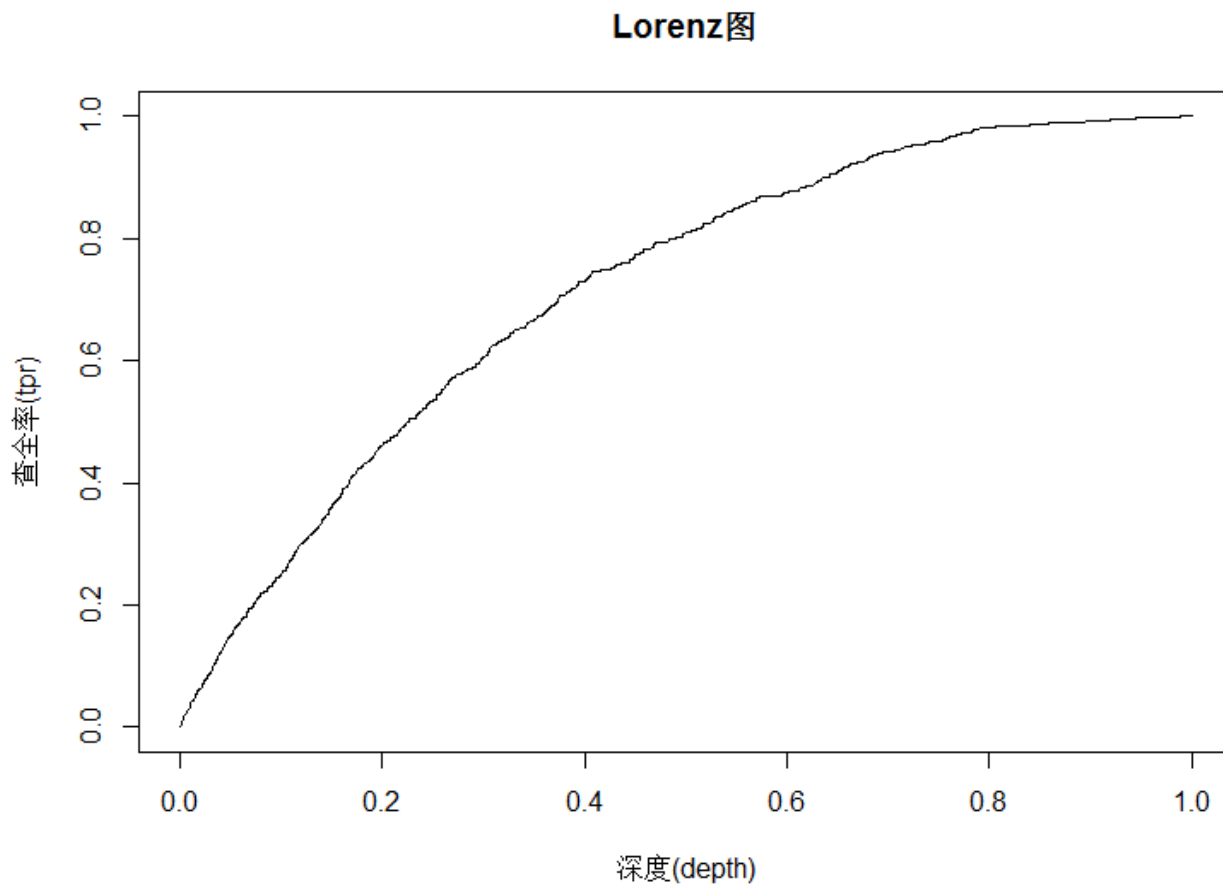


违约分值低处敏感

该模型在违约风险**低**人群中的预测能力较强，而在违约率**高**的部分较弱。有些业务需要做出这样的模型，比如VIP信用卡产品，业务需要低风险客户较高的信用额度，因此需要明确哪些客户的违约风险很低。

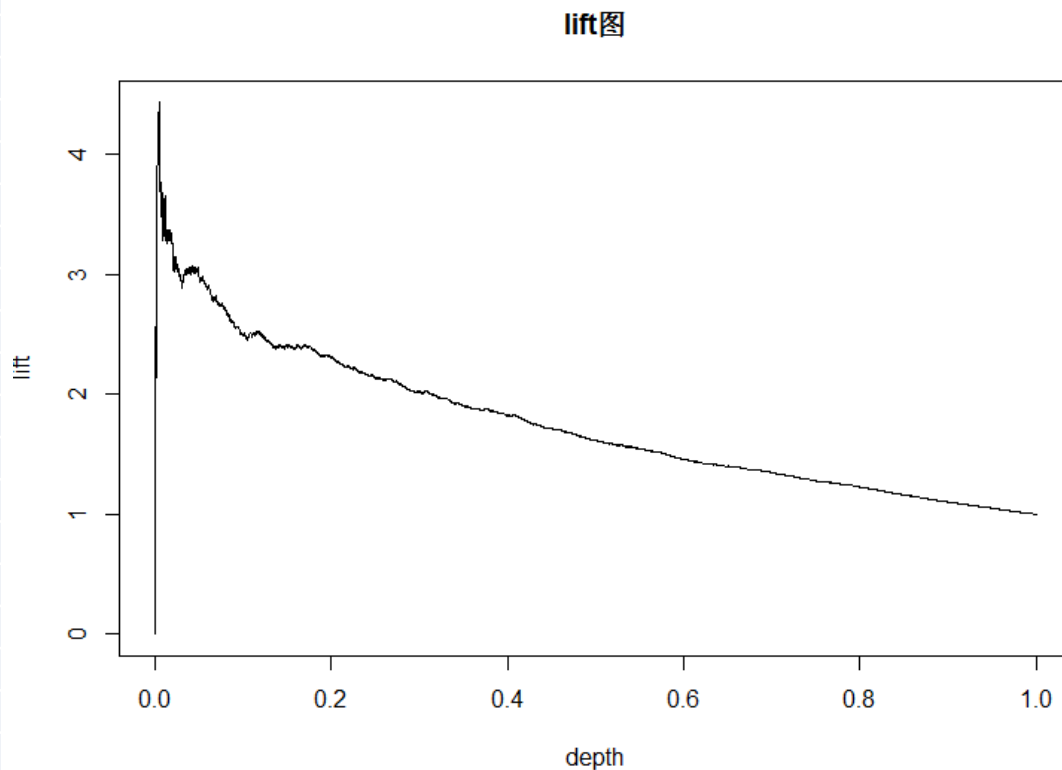
Gini 系数

深度	为真的组中的 累积百分比
0.0500	0.0900
0.1000	0.1800
0.1500	0.1800
0.2000	0.2700
0.2500	0.3600
0.3000	0.4500
0.3500	0.5500
0.4000	0.5500
0.4500	0.6400
0.5000	0.7300
0.5500	0.8200
0.6000	0.8200
0.6500	0.9100
0.7000	0.9100
0.7500	0.9100
0.8000	0.9100
0.8500	1.0000
0.9000	1.0000
0.9500	1.0000
1.0000	1.0000



累积提升度

深度	为真的组中的累积百分比	随机模型	累积提升度
0.0500	0.0900	0.0500	1.8000
0.1000	0.1800	0.1000	1.8000
0.1500	0.1800	0.1500	1.2000
0.2000	0.2700	0.2000	1.3500
0.2500	0.3600	0.2500	1.4400
0.3000	0.4500	0.3000	1.5000
0.3500	0.5500	0.3500	1.5714
0.4000	0.5500	0.4000	1.3750
0.4500	0.6400	0.4500	1.4222
0.5000	0.7300	0.5000	1.4600
0.5500	0.8200	0.5500	1.4909
0.6000	0.8200	0.6000	1.3667
0.6500	0.9100	0.6500	1.4000
0.7000	0.9100	0.7000	1.3000
0.7500	0.9100	0.7500	1.2133
0.8000	0.9100	0.8000	1.1375
0.8500	1.0000	0.8500	1.1765
0.9000	1.0000	0.9000	1.1111
0.9500	1.0000	0.9500	1.0526
1.0000	1.0000	1.0000	1.0000

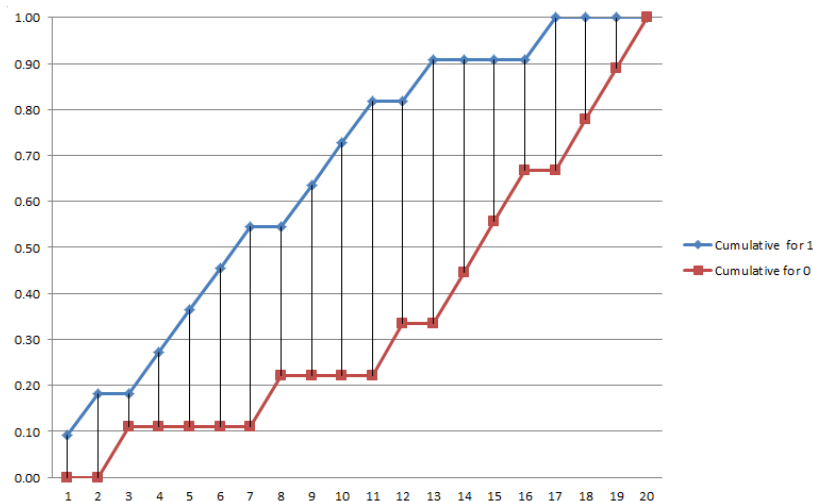


K-S统计量

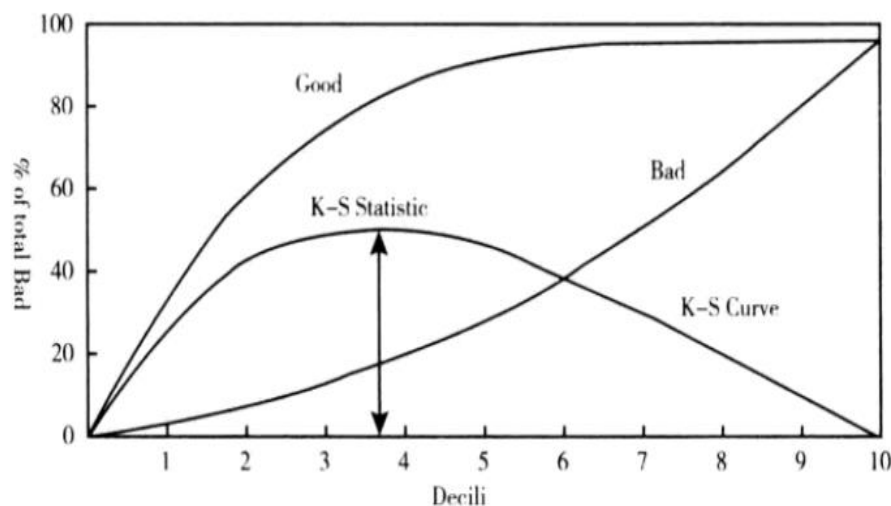
Predicted Probability	True Class	Cumulative for 1	Cumulative for 0	Difference*10
0.90	1	0.09	0.00	9
0.80	1	0.18	0.00	18
0.70	0	0.18	0.11	7
0.60	1	0.27	0.11	16
0.55	1	0.36	0.11	25
0.54	1	0.45	0.11	34
0.53	1	0.55	0.11	44
0.52	0	0.55	0.22	33
0.51	1	0.64	0.22	42
0.51	1	0.73	0.22	51
0.40	1	0.82	0.22	60
0.39	0	0.82	0.33	49
0.38	1	0.91	0.33	58
0.37	0	0.91	0.44	47
0.36	0	0.91	0.56	35
0.35	0	0.91	0.67	24
0.34	1	1.00	0.67	33
0.33	0	1.00	0.78	22
0.30	0	1.00	0.89	11
0.10	0	1.00	1.00	0

(K-S = 60)

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