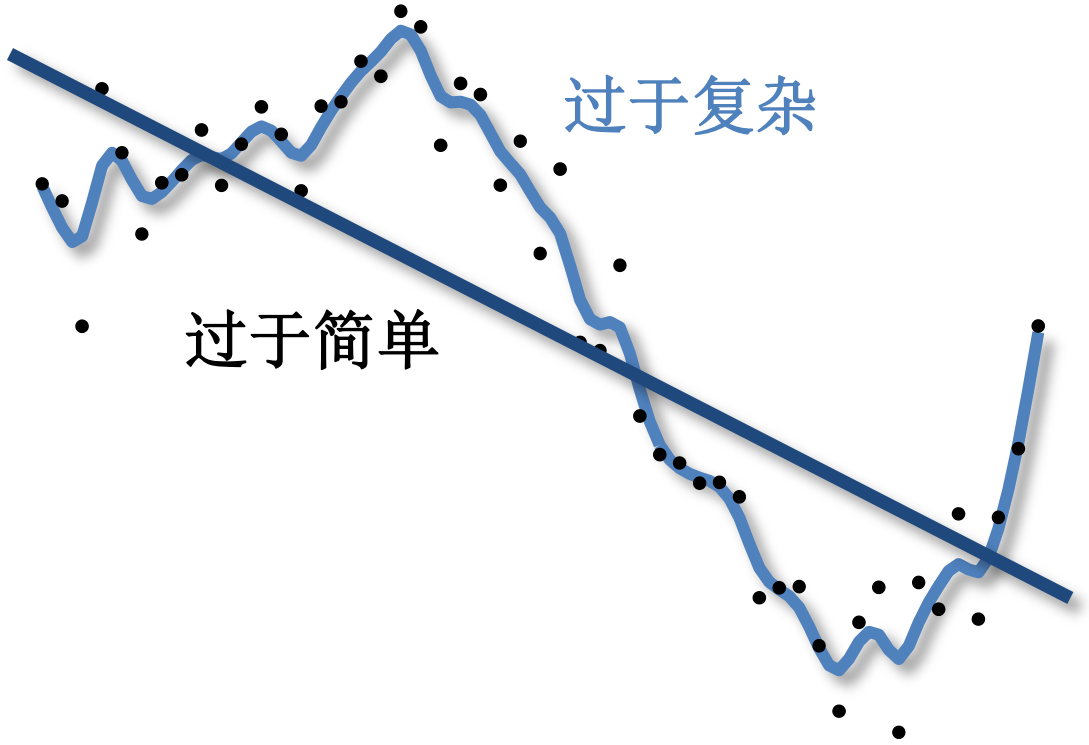
模型验证与模型选择

模型复杂度

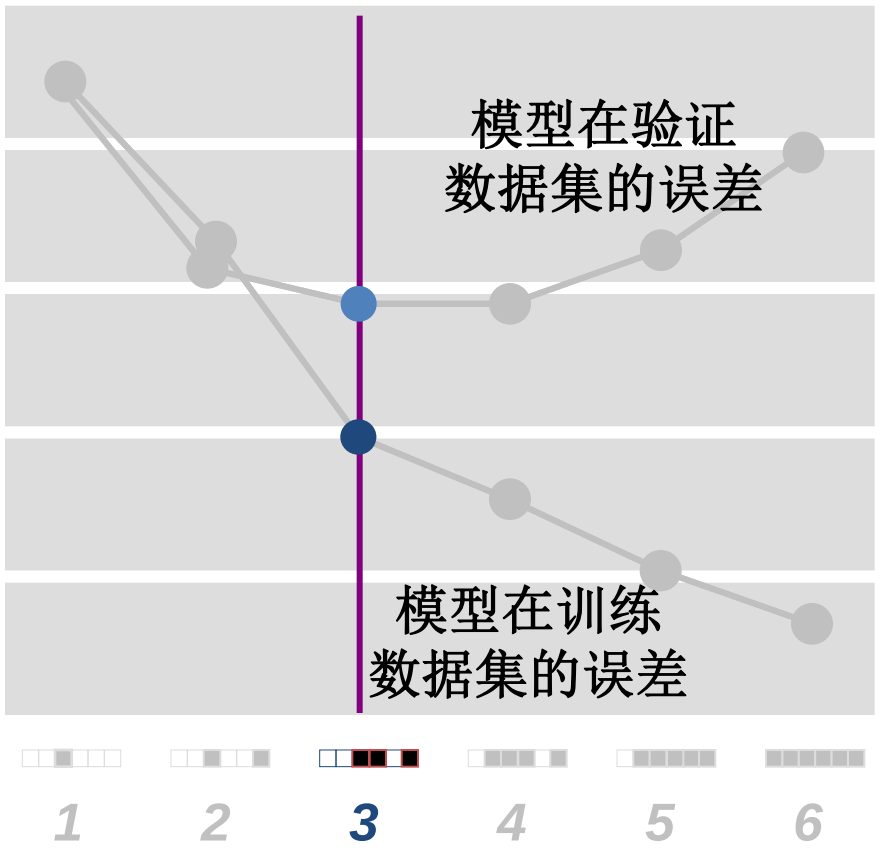
过于简单的模型预测精度不够，但是过于复杂的模型反映了数据中过多的噪音，该模型运用于其它数据集进行预测时，同样偏差较大。

这种过于复杂的模型称为过拟合。



模型复杂度

- 模型复杂度的选择根据模型的预测能力确定最优模型；
- 检验方法有两种：样本内检验，样本外检验；
- 样本内检验为根据一次抽样的结果，将原始数据集随机分为建模数据集、验证数据集和检验数据集；样本外验证通过滑动取数窗口，进行滚动建模和模型验证。 其中前一期数据用于训练和验证模型，当期数据用于检验。
- 一般建模中以样本内验证居多，右图是其示例。

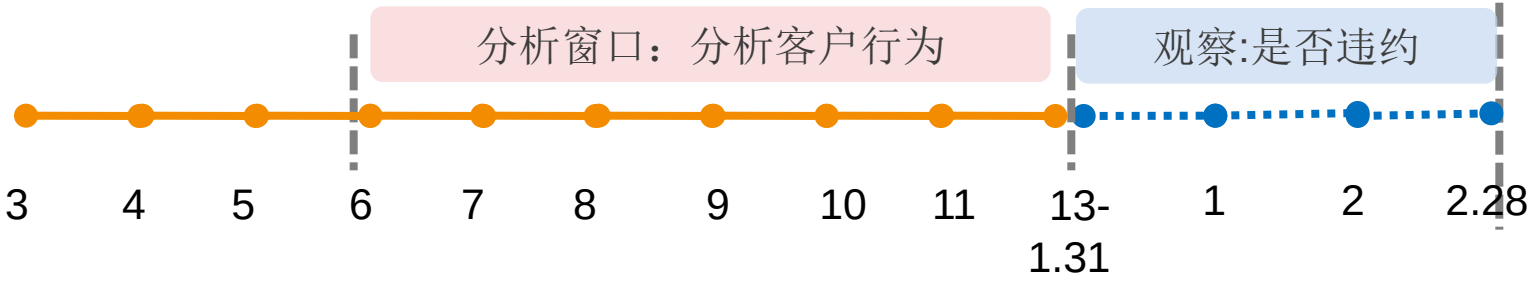


- 检验数据集用于检验模型的泛化能力。对于后剪枝的决策树模型是必须的，其他模型不是必须的。

模型检验

样本内检验：使用训练集同期的数据

样本外检验：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）	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
估计（Estimates）（省略	误差平方均值 SBC/可能性

对评估数据集进行处理

- 评估数据集同样需要进行数据清洗、缺失值填补、分类变量WOE转换等操作；
- 在缺失值填补等操作时，需要使用使用训练数据集的统计量，而不是验证数据集的统计量。

决策模型主要指标

TP——将正类预测为正类数；

FN——将正类预测为负类数；

FP——将负类预测为正类数；

TN——将负类预测为负类数.

精确率定义为

$$P=\frac{TP}{TP+FP}$$

召回率定义为

$$R=\frac{TP}{TP+FN}$$

此外，还有 F_1 值，是精确率和召回率的调和均值，即

$$\frac{2}{F_1}=\frac{1}{P}+\frac{1}{R}$$

$$F_1=\frac{2TP}{2TP+FP+FN}$$

混淆矩阵（Confusion Matrix）

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。

		打分值(Predicted Class)		
		反应（预测=1）	未反应（预测=0）	合计
真实结果 (Actual Class)	呈现信号（真实=1）	A（击中, True Positive）	B（漏报, False Negative）	A + B, Actual Positive
	未呈现信号（真实=0）	C（虚报, False Positive）	D（正确否定, True Negative）	C + D, Actual Negative
合计		A + C, Predicted Positive	B + D, Predicted Negative	A + B + C + D

灵敏度= A /(A + B)
特异度 = D /(C + D)

混淆矩阵

预测 / 实际	1	0
1	d True Positive	c False Negative
0	b False Positive	a True Nagative

3-1.准确率或者误分类率　正确率=（a+d)/(a+b+c+d)
2. 覆盖率（**recall**，**Sensitivity**） d/(c+d）
3.命中率(**Precision,PV+**) d/(b+d)
4.负例的覆盖率(**Specificity)** a/(a+b)
5.PV-　a/(a+c)

混淆矩阵（Confusion Matrix）

Predicted Probability	True Class
0.90	1
0.80	1
0.70	0
0.60	1
0.55	1
0.54	1
0.53	1
0.52	0
0.51	1
0.51	1
0.40	1
0.39	0
0.38	1
0.37	0
0.36	0
0.35	0
0.34	1
0.33	0
0.30	0
0.10	0

Confusion Matrix		True Results	
		0	1
Predicted Results	0	7	3
	1	2	8

Accuracy = 0.75 & Misclassification = 0.25

Precision = 0.8 & Recall = $\frac{8}{11}$ = 0.73

F−Measure = 0.76

Sensitivity = True Positive Rate = Recall

Specificity = True Negative Rate = $\frac{7}{9}$ = 0.78

混淆矩阵（Confusion Matrix）

Confusion Matrix		True Results	
		0(Secondary)	1(Primary)
Predicted Results	0(Secondary)	TN(真陰性)	FN(偽陰姓)
	1(Primary)	FP(偽陽性)	TP(真陽性)

$$\text{Accuracy} = \frac{TN+TP}{TN+FN+FP+TP} \quad \text{Misclassification} = \frac{FN+FP}{TN+FN+FP+TP}$$

$$\text{Precision} = \frac{TP}{FP+TP} \quad \text{Recall} = \frac{TP}{FN+TP}$$

$$\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Sensitivity} = \text{True Positive Rate} = \text{Recall}$$

$$\text{Specificity} = \text{True Negative Rate} = \frac{TN}{TN+FP}$$

混淆矩阵（Confusion Matrix）

Predicted Probability	True Class
0.90	1
0.80	1
0.70	0
0.60	1
0.55	1
0.54	1
0.53	1
0.52	0
0.51	1
0.51	1
0.40	1
0.39	0
0.38	1
0.37	0
0.36	0
0.35	0
0.34	1
0.33	0
0.30	0
0.10	0

Confusion Matrix		True Results	
		0	1
Predicted Results	0	6	1
	1	3	10

Accuracy = 0.8 & Misclassification = 0.2

Precision = 0.77 & Recall = $\frac{10}{11}$ = 0.91

F–Measure = 0.83

Sensitivity = True Positive Rate = Recall

Specificity = True Negative Rate = $\frac{6}{9}$ = 0.67

排序模型主要指标

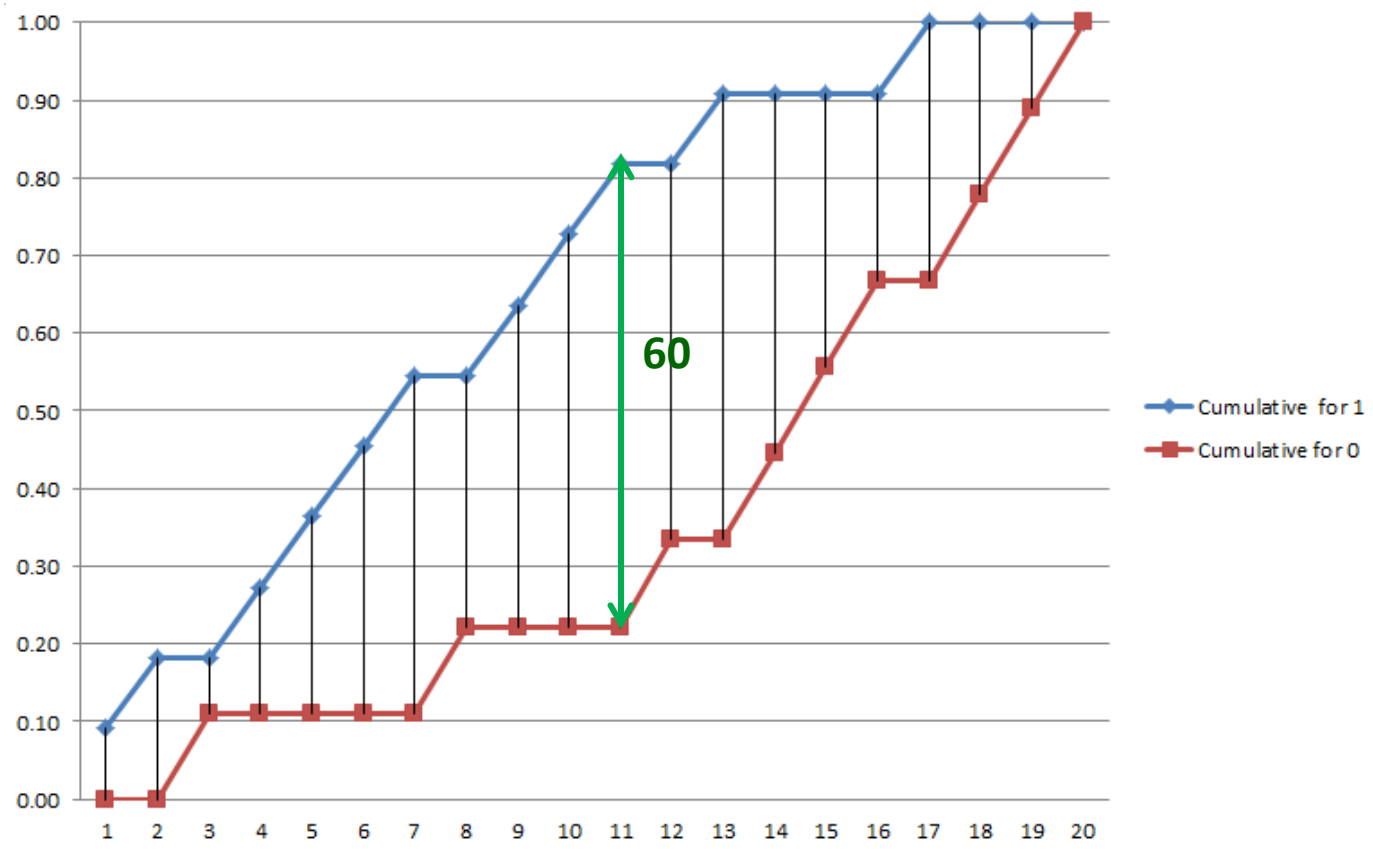
- K-S
- ROC
- Gini

K-S统计量

Predicted Probability	True Class	Cumulative for 1	Cumulative for 0	Difference*100
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	3-1.00	0.67	33
0.33	0	3-1.00	0.78	22
0.30	0	3-1.00	0.89	11
0.10	0	3-1.00	3-1.00	0

(K-S = 60)

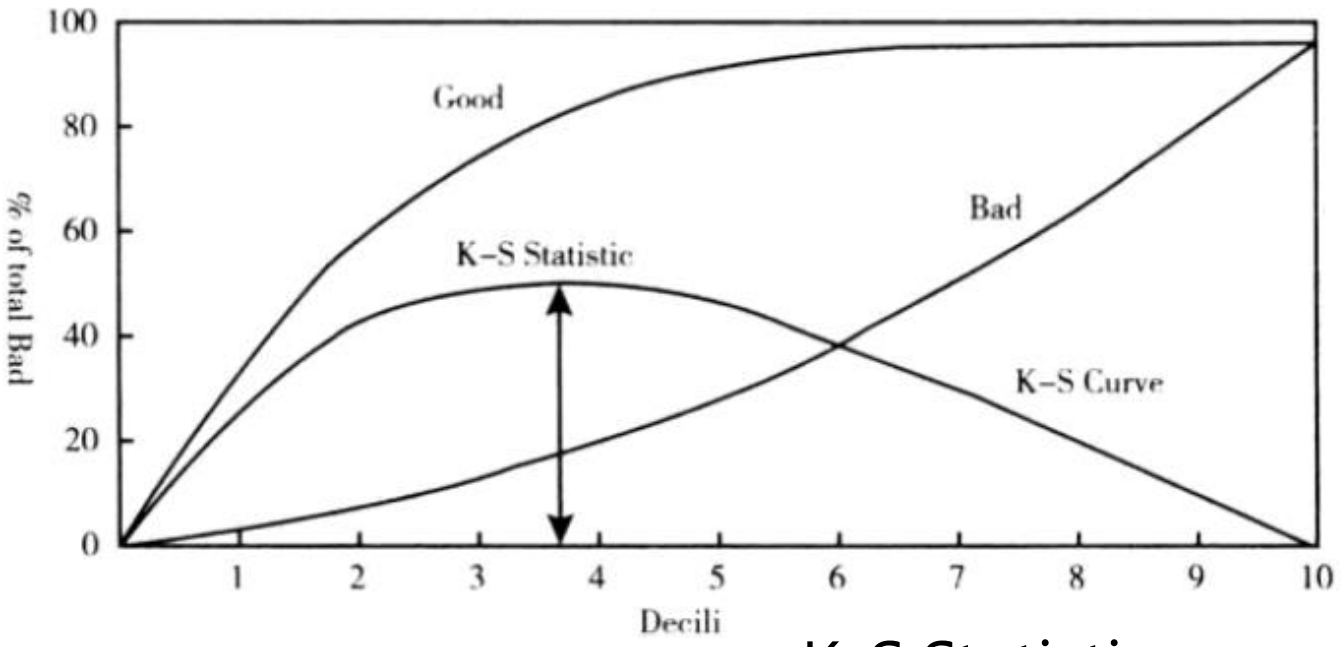
K-S Chart


1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
60
Cumulative for 1
Cumulative for 0

K-S Chart

- K-S (Kolmogorov-Smirnov) chart measures performance of classification models
- K-S is a measure of the degree of separation between the positive and negative distributions
  - The K-S is 100 if the scores partition the population into two separate groups in which one group contains all the positives and the other all the negatives
  - If the model cannot differentiate between positives and negatives, then the K-S would be 0 (the model selects cases randomly)
  - The higher the value the better the model is at separating the positive from negative cases

K-S曲线



- K-S Statistics
  - 小於20
    - 此模型無鑑別力
  - 20~40之間
    - 此模型勉強接受
  - 41~50之間
    - 此模型有好的區別能力

- K-S Statistics
  - 51~60之間
    - 此模型有很好的區別能力
  - 61~75之間
    - 此模型有非常好的區別能力
  - 大於75
    - 此模型異常，可能有問題

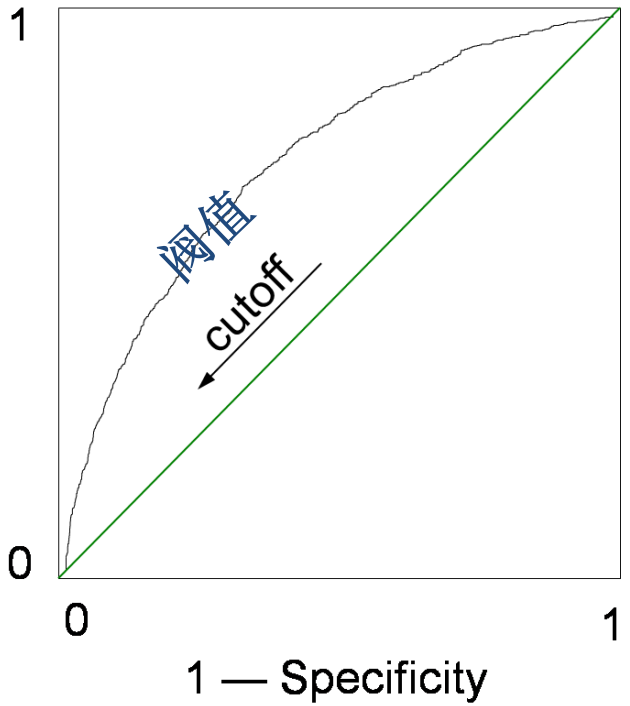
ROC图形

	阈值	敏感度	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

阀值下降

灵敏度

Sensitivity

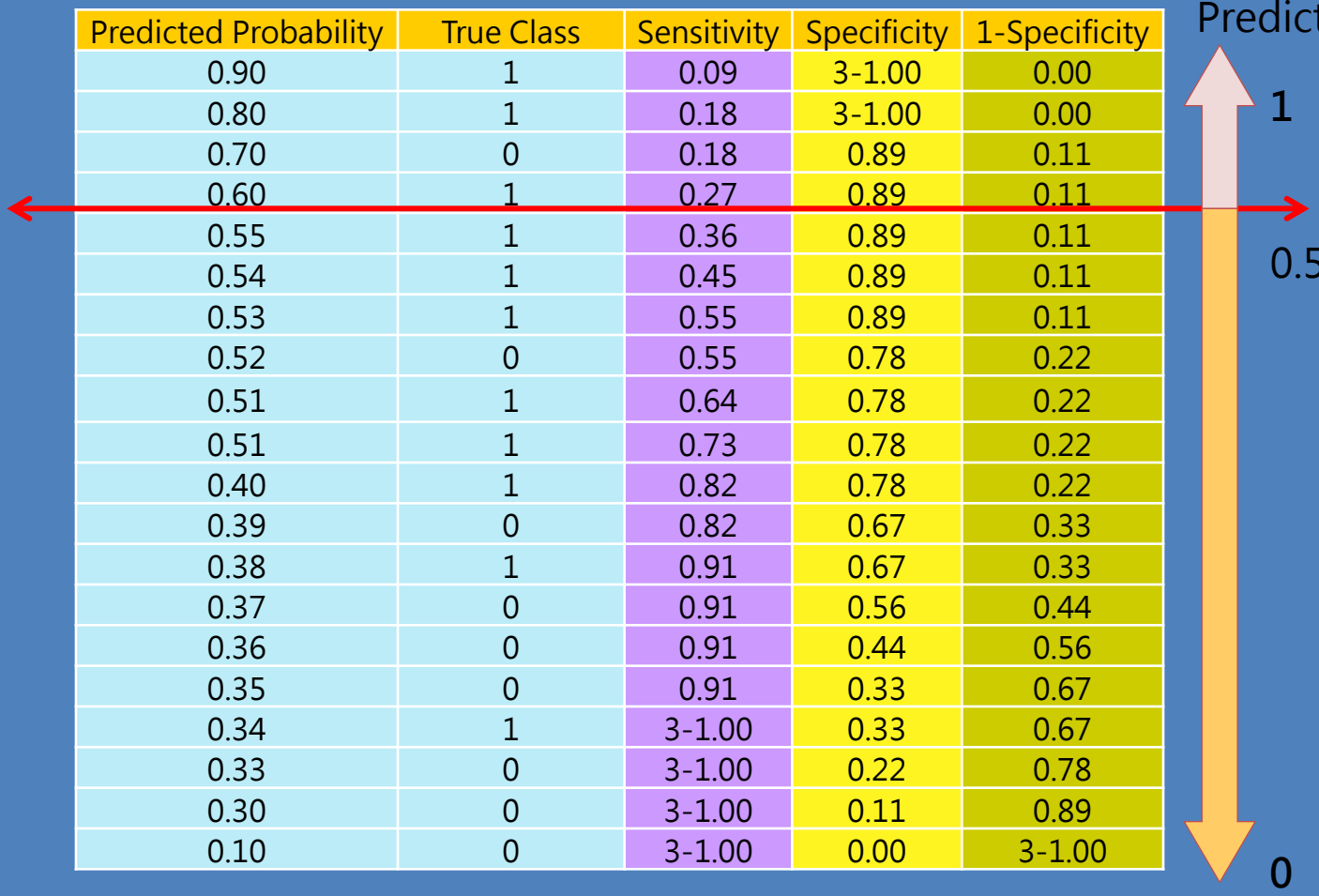


1-特异度

随着阀值的下降，灵敏度在升高，特异度在降低。

Statistical Graphics – ROC Chart

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



CM		True	
		0	1
P	0	8	8
	1	1	3

Accuracy = 0.55
Precision = 0.75
Recall = 0.27

Sensitivity = Recall

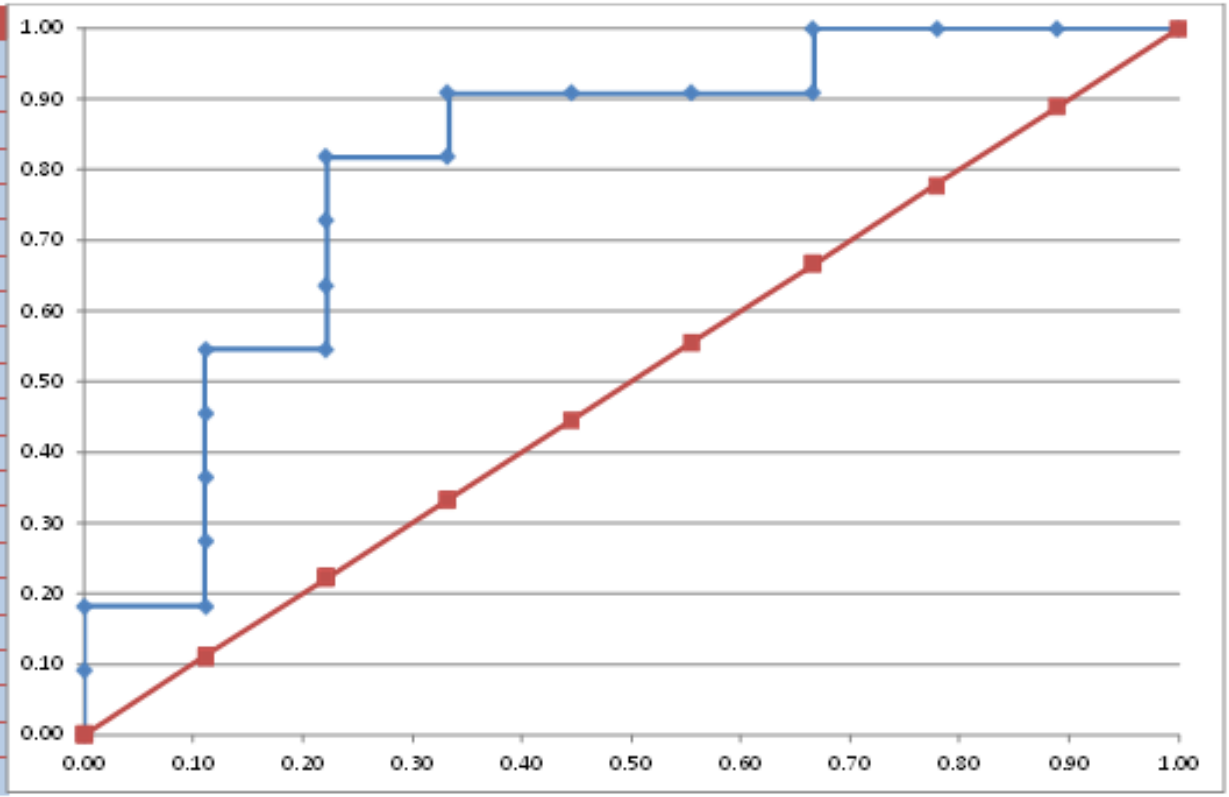
Specificity
= True Negative Rate
= 0.89

Statistical Graphics – ROC Chart

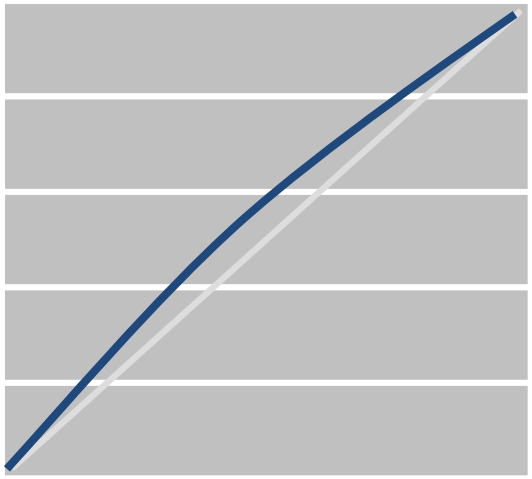
Predicted Probability	True Class	Cumulative for 1	Cumulative for 0	Sensitivity	Specificity	1-Specificity
0.90	1	0.09	0.00	0.09	3-1.00	0.00
0.80	1	0.18	0.00	0.18	3-1.00	0.00
0.70	0	0.18	0.11	0.18	0.89	0.11
0.60	1	0.27	0.11	0.27	0.89	0.11
0.55	1	0.36	0.11	0.36	0.89	0.11
0.54	1	0.45	0.11	0.45	0.89	0.11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

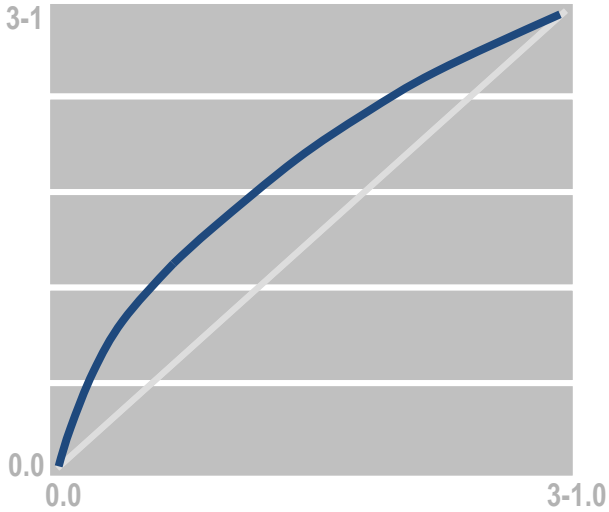
Statistical Graphics – ROC Chart

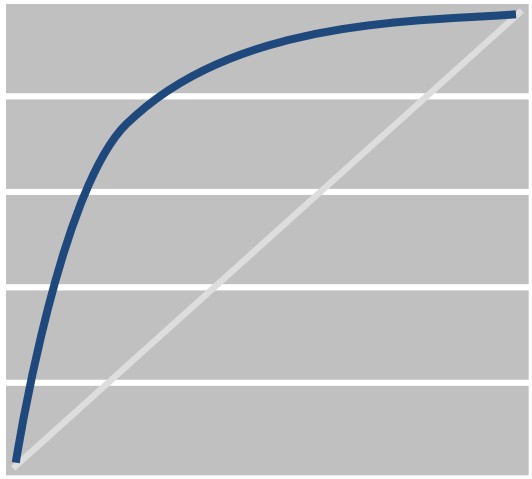
1-Specificity	Random	Sensitivity
0.00	0.00	0.00
0.00	0.00	0.09
0.00	0.00	0.18
0.11	0.11	0.18
0.11	0.11	0.27
0.11	0.11	0.36
0.11	0.11	0.45
0.11	0.11	0.55
0.22	0.22	0.55
0.22	0.22	0.64
0.22	0.22	0.73
0.22	0.22	0.82
0.33	0.33	0.82
0.33	0.33	0.91
0.44	0.44	0.91
0.56	0.56	0.91
0.67	0.67	0.91
0.67	0.67	1.00
0.78	0.78	1.00
0.89	0.89	1.00
1.00	1.00	1.00



ROC 图形





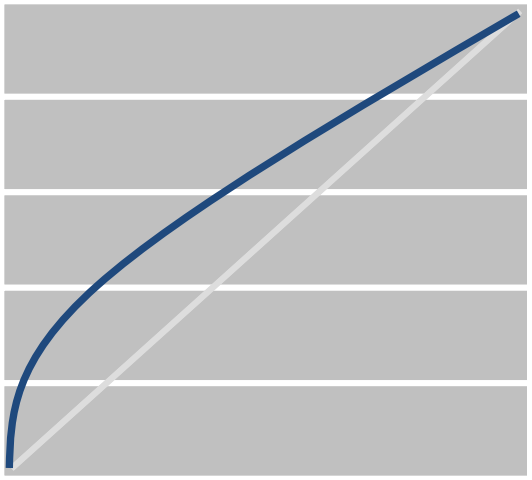


弱的模型

强的模型

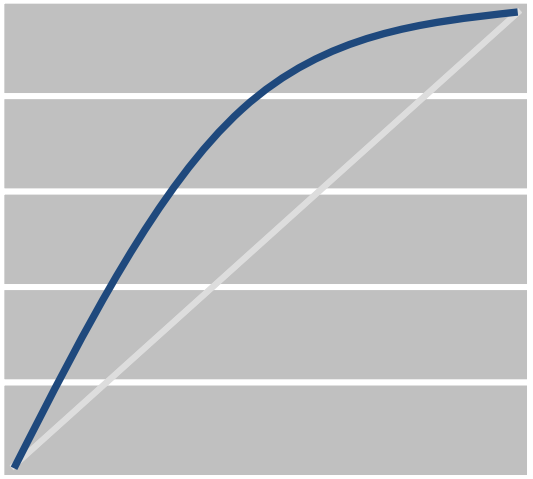
ROC曲线结果的取值在[0.5,1]。
一般来说，
[0.5,0.7)表示效果较低；
[0.7,0.85) 表示效果一般；
[0.85,0.95)表示效果良好；
[0.95,0.1]社会科学建模中不大可能出现。
注意：
①有时ROC曲线可能会落入对角线以下，这时需检查检验方向与状态值的对应关系
②如果某ROC曲线在对角线两边均有分布，需检查数据或专业背景。

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。



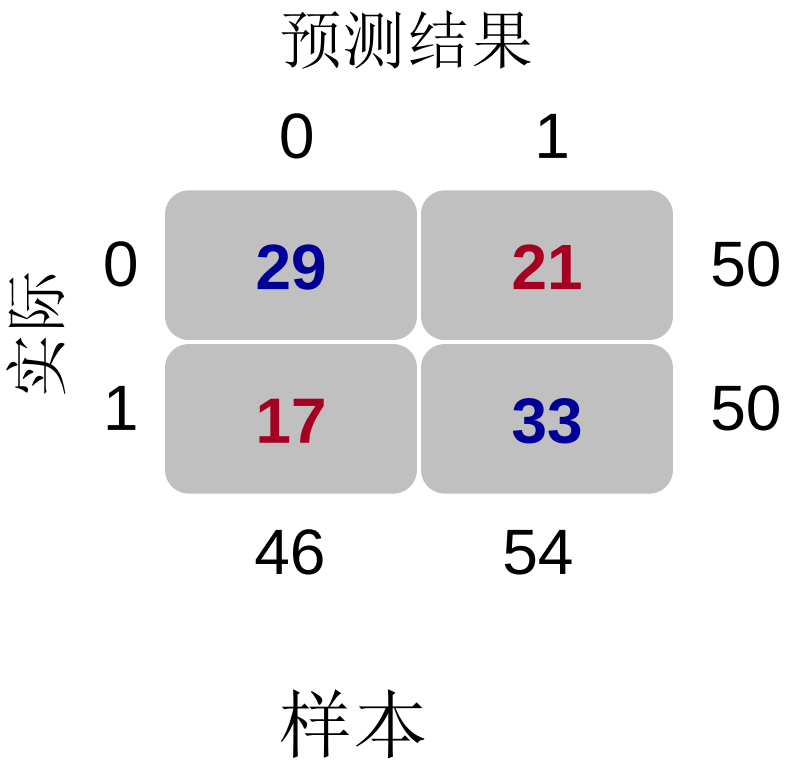
违约分值低处敏感

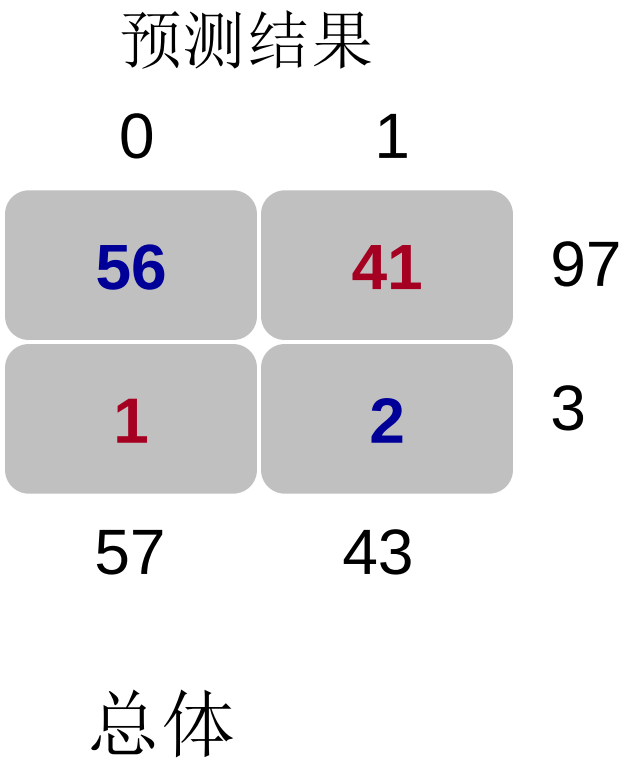
该模型在违约风险**低**人群中的预测能力较强，而在违约率**高**的部分较弱。有些业务需要做出这样的模型，比如VIP信用卡产品，业务需要低风险客户较高的信用额度，因此需要明确哪些客户的违约风险很低。

Area Under the Curve (AUC)

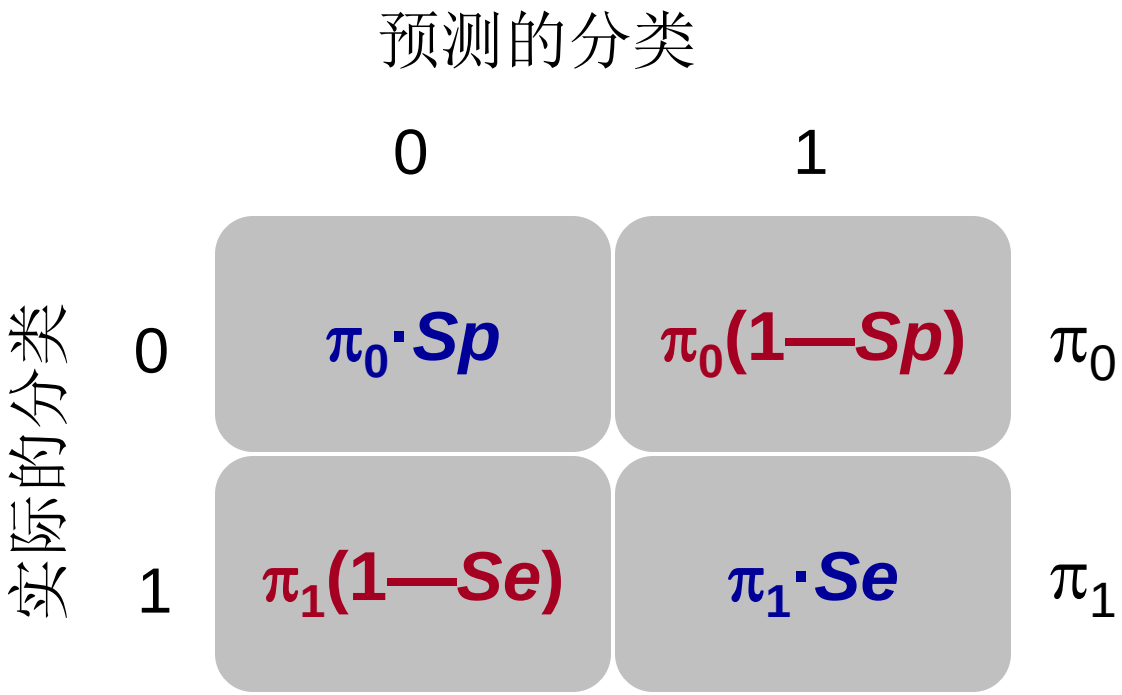
- Area under ROC curve is often used as a measure of quality of the classification models
- A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1
- In practice, most of the classification models have an AUC between 0.5 and 1

如果使用过抽样

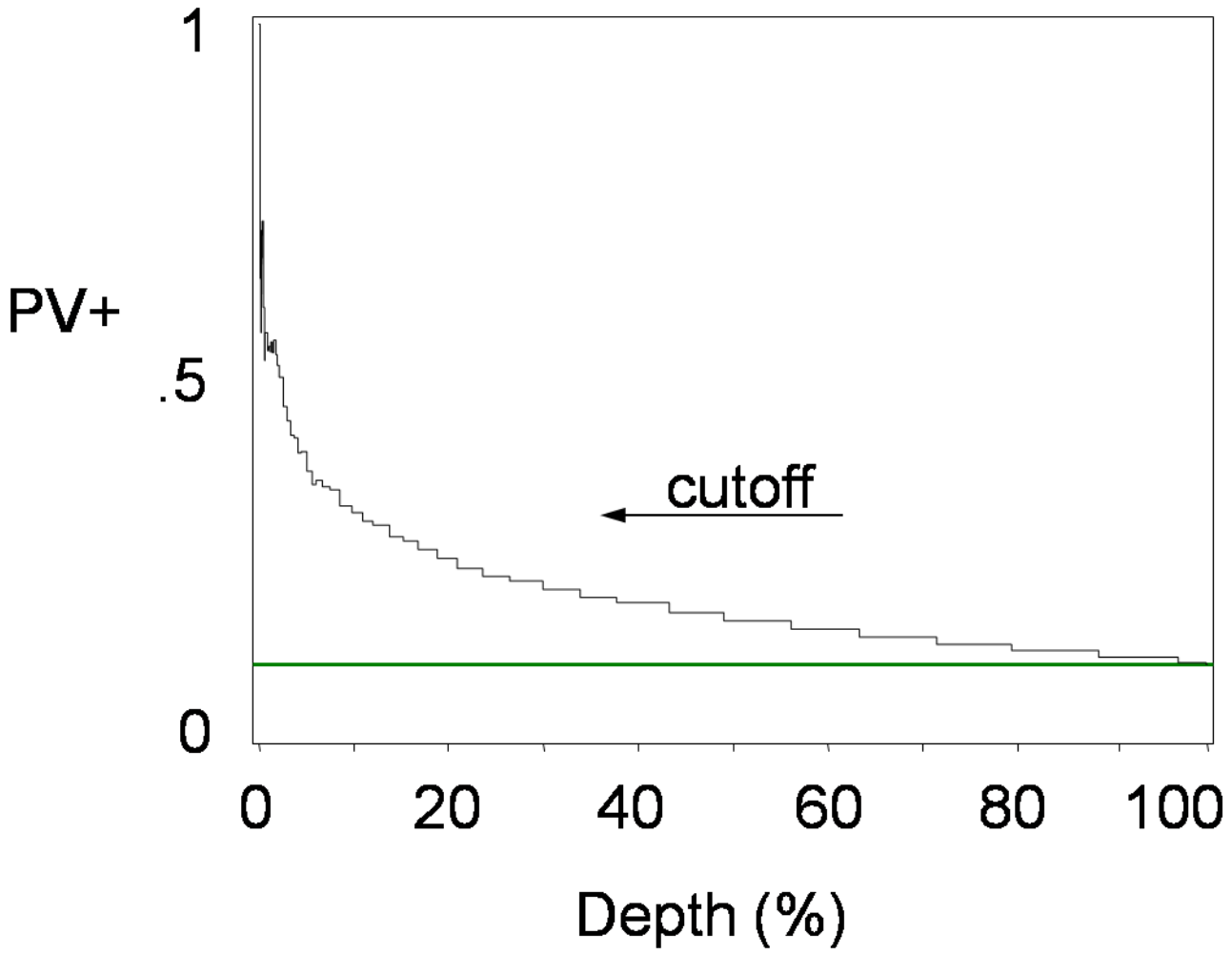




过抽样调整



Gains Chart

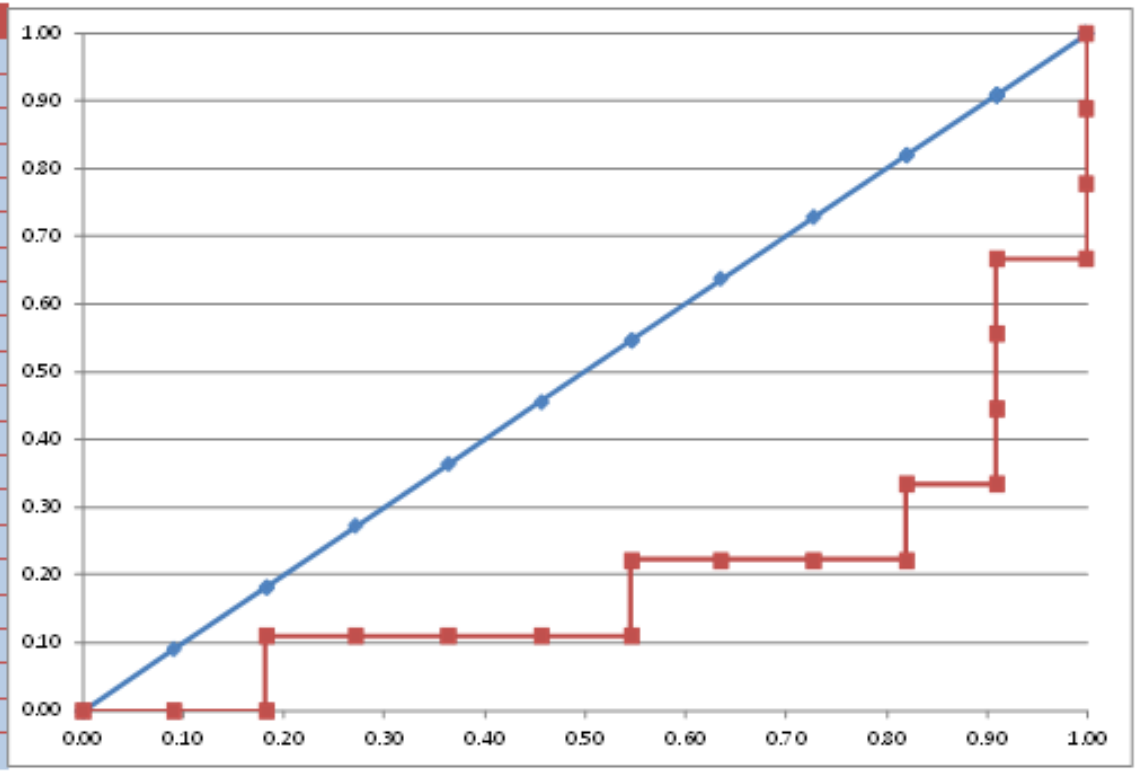


Gini Coefficient

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

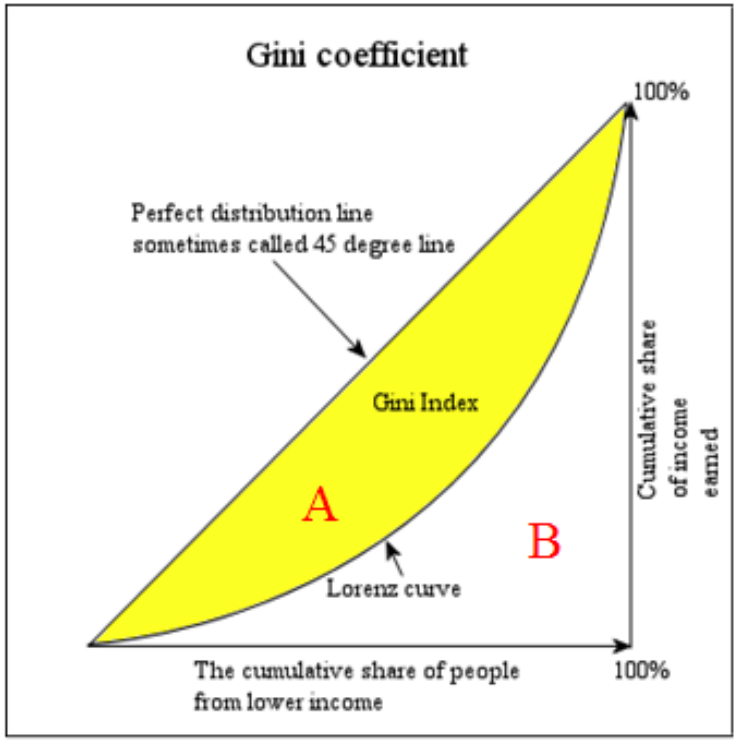
Gini Coefficient

Cumulative for 1	Random	Cumulative for 0
0.00	0.00	0.00
0.09	0.09	0.00
0.18	0.18	0.00
0.18	0.18	0.11
0.27	0.27	0.11
0.36	0.36	0.11
0.45	0.45	0.11
0.55	0.55	0.11
0.55	0.55	0.22
0.64	0.64	0.22
0.73	0.73	0.22
0.82	0.82	0.22
0.82	0.82	0.33
0.91	0.91	0.33
0.91	0.91	0.44
0.91	0.91	0.56
0.91	0.91	0.67
1.00	1.00	0.67
1.00	1.00	0.78
1.00	1.00	0.89
1.00	1.00	1.00



Gini Coefficient vs. ROC Index

- The Gini coefficient is calculated as a ratio of the areas on the Lorenz curve diagram
  - A : the area between the line of perfect equality and Lorenz curve
  - B : the area underneath the Lorenz curve
  - Gini coefficient is A/(A+B)



ROC Index = A + 0.5
Gini Coefficient = A / (A+B) = 2A = 2 * (ROC Index – 0.5)

使用汇总统计比较模型

- 演示说明模型的比较工具的使用，从连接的建模节点收集评估信息，使你更容比较易模型的性能。