模型选择与评估

- 模型选择方法：
  - Hold Out
  - Cross Validation
  - Bootstrap

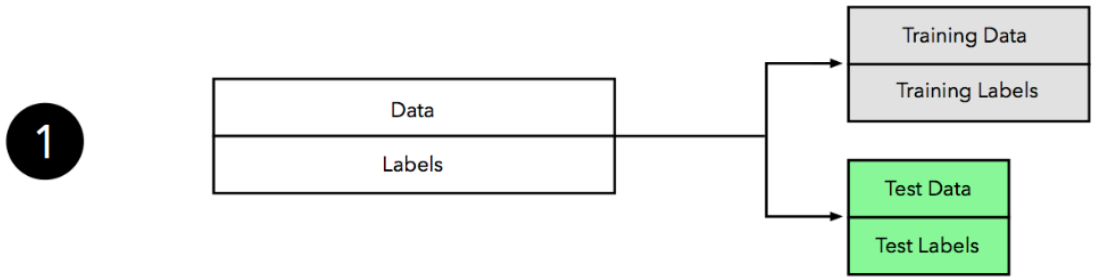
- 模型评估方法
  - Confusion Matrix
  - ROC曲线 与 AUC
  - Lift(提升)和Gain(增益)

- holdout cross-validation 以及 k-fold cross-validation 都是交叉验证的方法。

- 评价模型在新的数据集上的性能。

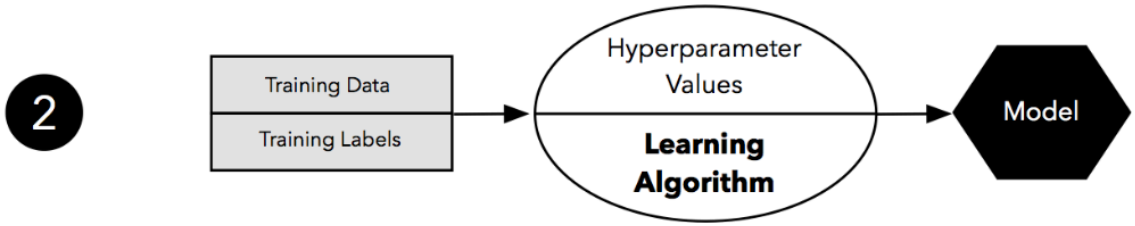
- 平衡过拟合与欠拟合。

- holdout cross-validation 过程



划分训练集，一般训练集大概在2/3左右。

- holdout 过程

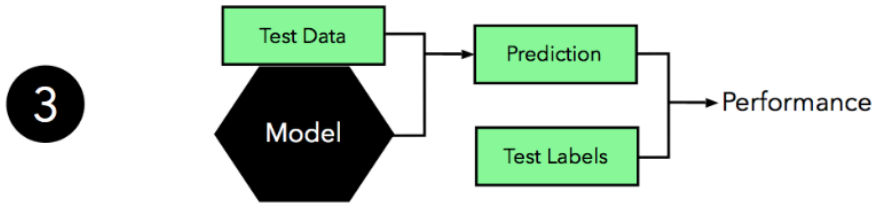


调节超参数。

超参数：机器学习模型里面的框架参数，如聚类个数，正则化系数，KNN中的K的选取，一般人工手动调节。

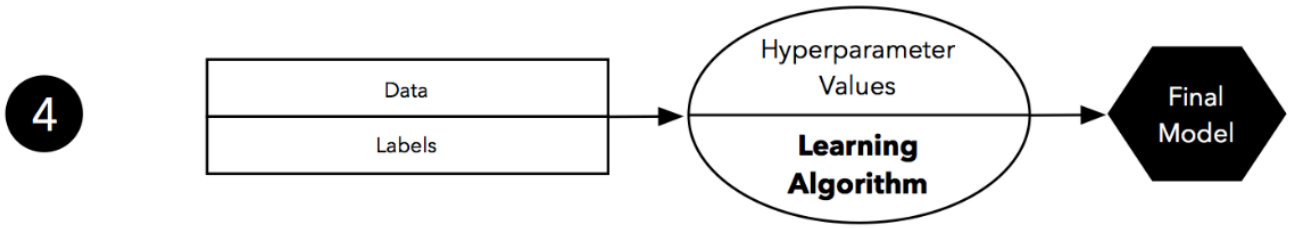
学习参数：由数据学习到的权重，不需要手动调节，由数据决定。

- holdout cross-validation 过程



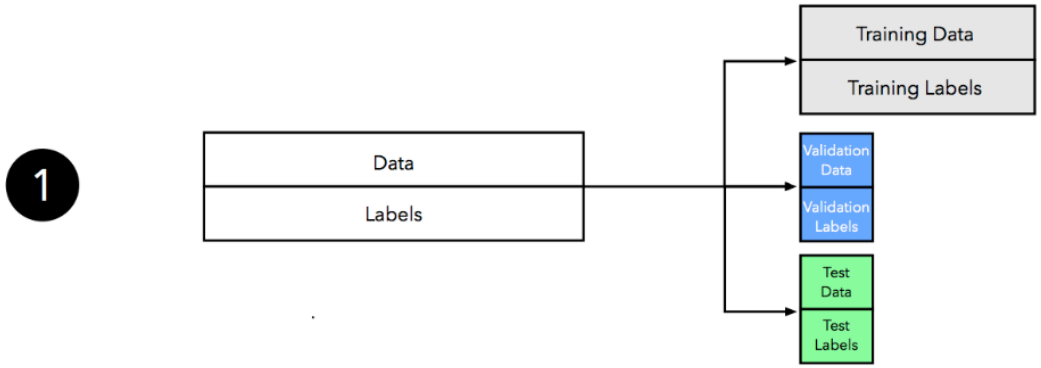
训练好的模型需要用测试集做验证。通常情况下，还需要对训练集进行评估。

- holdout cross-validation 过程



超参数确定了，还有一部分没有参与训练，利用全量数据集训练模型。得到最终的结果。

- 上述方法存在一定的偶然性，并且不能监督过程。
- 引出 three-way holdout method

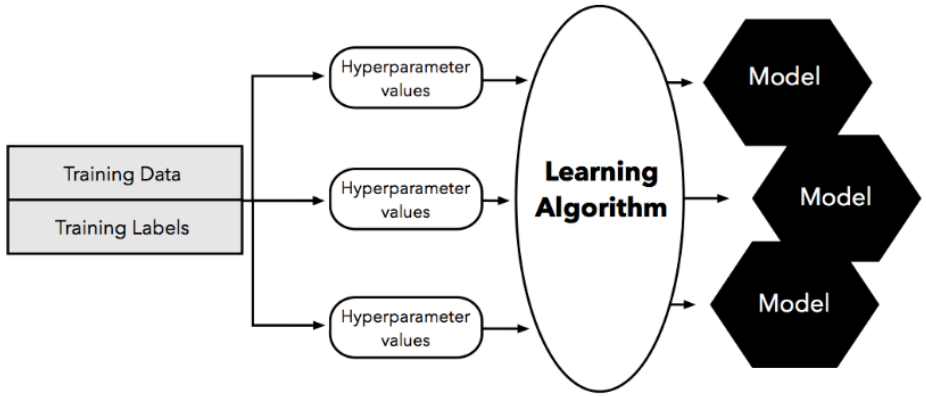


分为训练集、验证集、测试集。

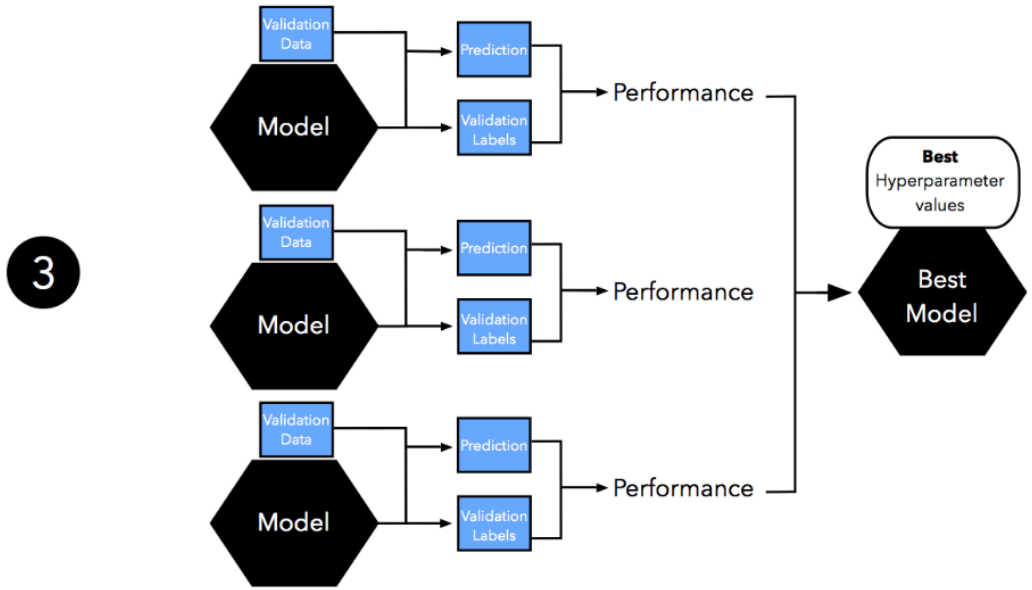
测试集的作用：参与监督过程和调参指导。

- 采用不同的超参数在训练集上训练同一模型。

2

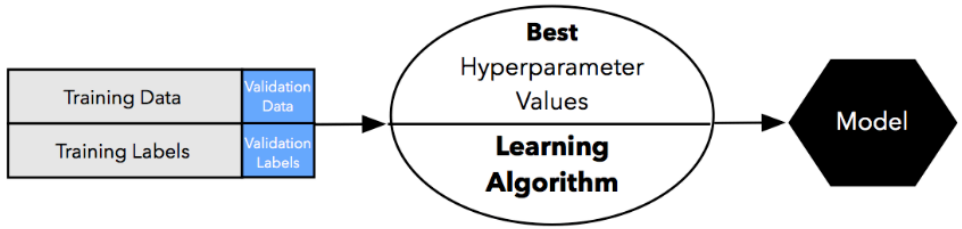


- 训练过程中利用验证集挑选最优超参数，选择最好的一个模型。



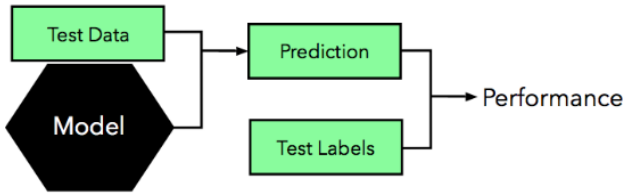
- 挑选出了最优的超参数，那么将训练集和验证集合在一起训练。

4

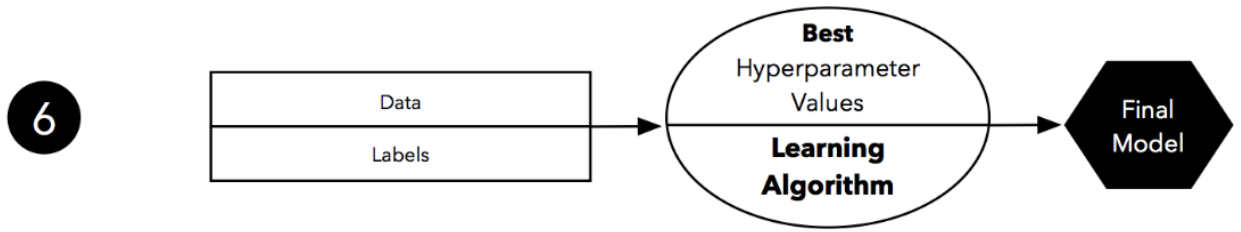


- 在测试集上的表现结果。

5



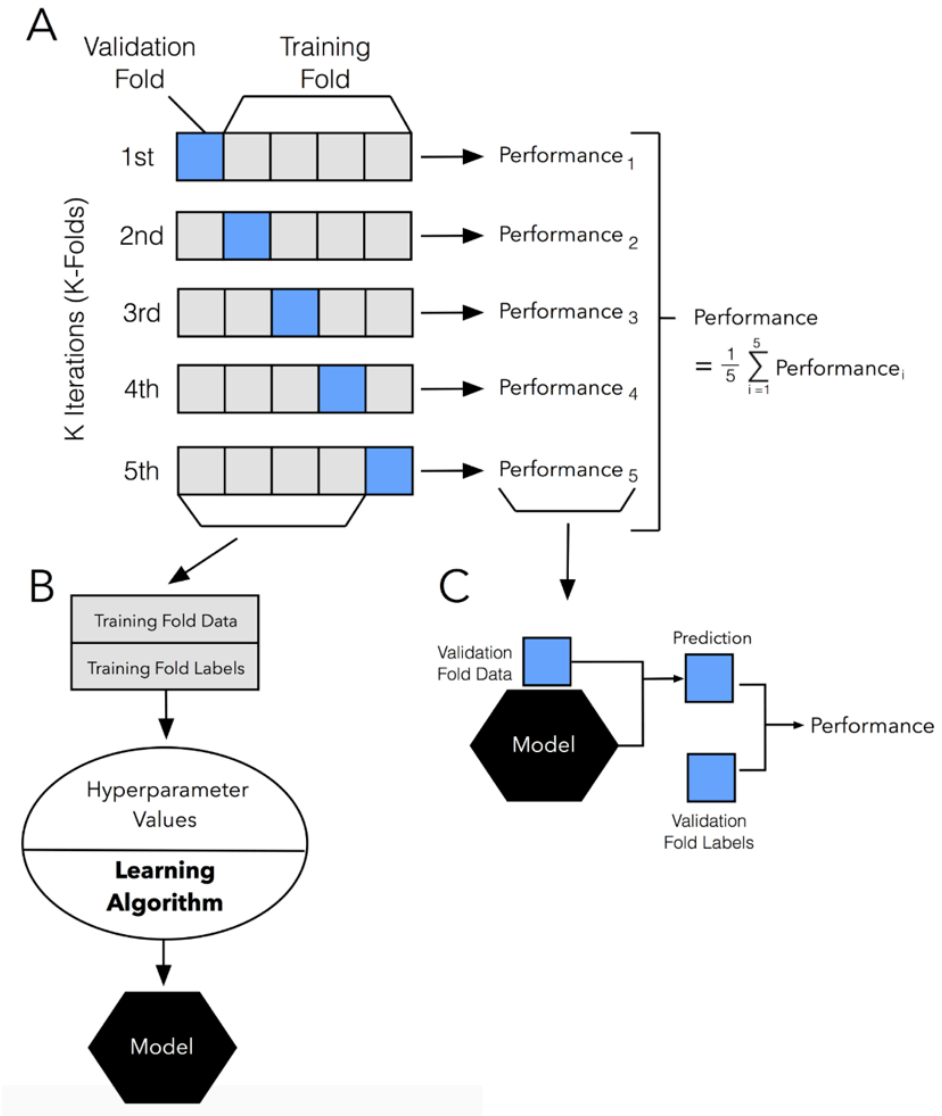
- 所有的训练数据放在一起，拿到模型中训练。



- cross-validation

K-fold 交叉验证

优点：利用全部数据集
更加鲁棒



- Bootstrap
  - 自助法
  - 一种重采样（Resampling）技术
  - 集成学习中会用到它的思想

有放回的抽样；

是否有的样本压根没有采样到

- 这个概率是可以计算的：

- 有n个样本，每个样本的取到的概率是1/n

$$P(\text{not chosen}) = \left(1 - \frac{1}{n}\right)^n$$

如果n取∞，则极限为1/e ,~0.368

Random Forest

通过常用极限推导而来 $\lim_{x \to 0}(1+x)^{\frac{1}{x}} = e$

- 天然的划分训练集和测试集。
- 实际上，Random Forest 模型验证就是采用这种办法。

模型评估

- Confusion Matrix
  - 对于二分类问题，可将样例根据其真实类别与学习期预测类别的组合划分为真正例(True Positive),假正例(False Positive)，真反例(True Negative)，假反例(False Negative)四种情形，四种情形组成的混淆矩阵如下：

真实情况	预测结果	
	正例	反例
正例	TP	FN
反例	FP	TN

模型评估

真实情况	预测结果	
	正例	反例
正例	TP	FN
反例	FP	TN

准确率(Accuracy)
Accuracy=(TP+TN) / (TP+FP+TN+FN)

精准率(Precision)
Precision=TP / (TP+FP)

召回率(Recall)
Recall=TP / (TP+FN)

F值：F-score是Precision和Recall加权调和平均数，并假设两者一样重要

F1 Score=(2*Recall*Precision) / (Recall+Precision)

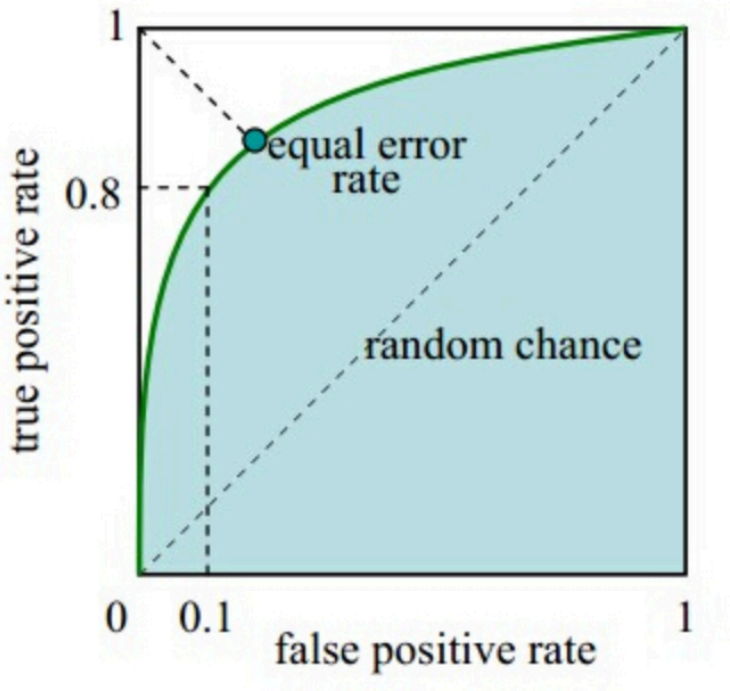
模型评估

- 某池塘有1400条鲤鱼，300只虾，300只鳖。现在以捕鲤鱼为目的。撒一大网，逮着了700条鲤鱼，200只虾，100只鳖。那么，精确率和召回率分别为多少？

- 精确率 = 700/(700+200+100)
- 召回率 = 700/1400

模型评估

- ROC 与AUC
- ROC
- AUC



ROC全称是“受试者工作特征”（Receiver Operating Characteristic）。ROC曲线的面积就是AUC（Area Under the Curve）。AUC用于衡量“二分类问题”机器学习算法性能（泛化能力）。

模型评估

- 如何画Roc

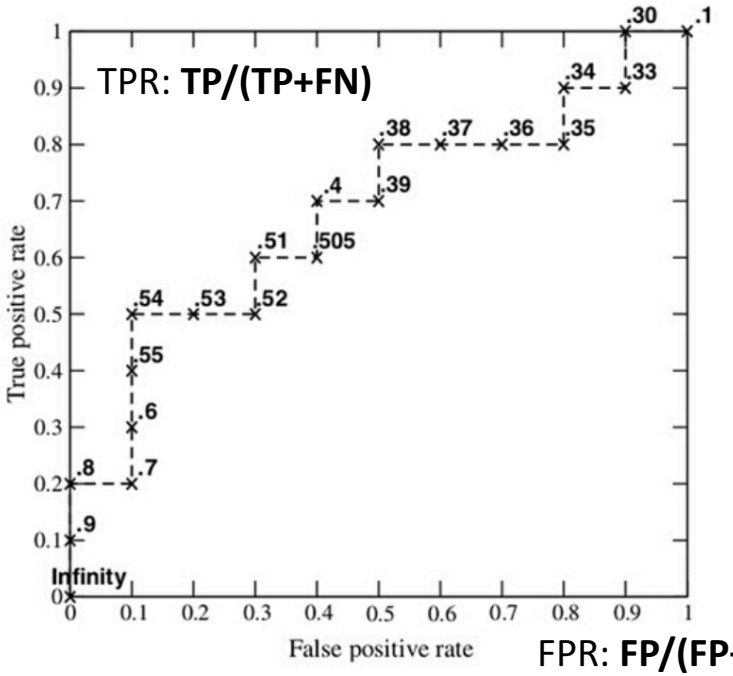
- 真正类率(True Postive Rate)
- TPR: **TP/(TP+FN)**
- 负正类率(False Postive Rate)
- FPR: **FP/(FP+TN)**

真实情况	预测结果	
	正例	反例
正例	TP	FN
反例	FP	TN

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

模型评估

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

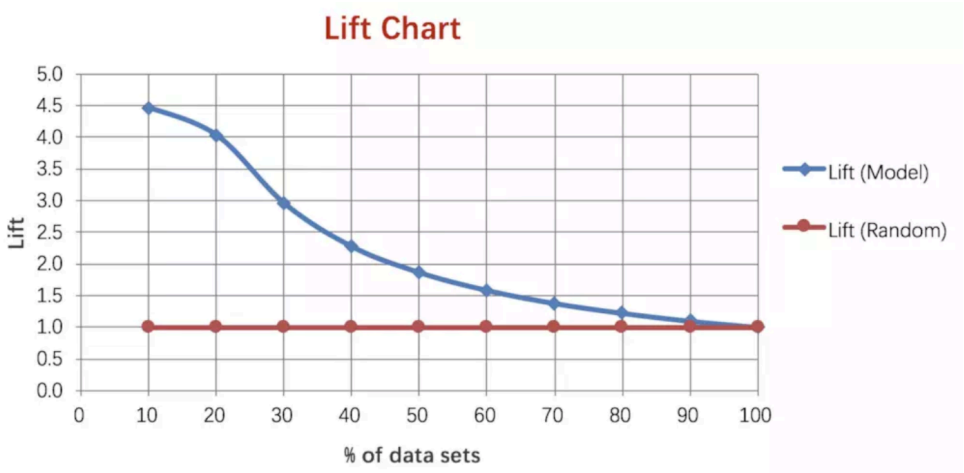


FPR: **FP/(FP+TN)**

真实情况	预测结果	
	正例	反例
正例	TP	FN
反例	FP	TN

真实情况	预测结果	
	正例	反例
正例	TP	FN
反例	FP	TN

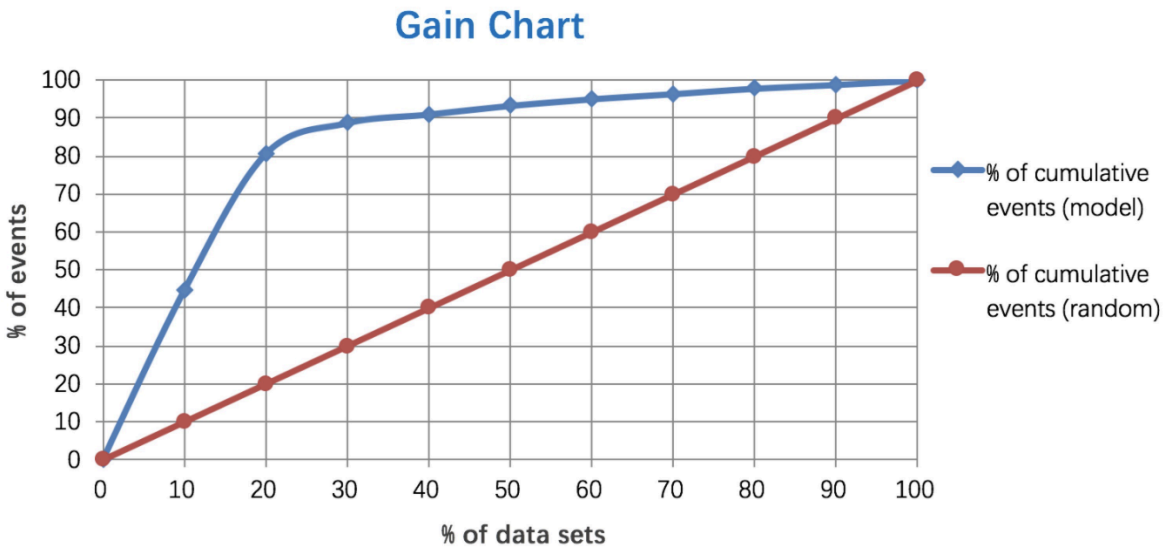
- Lift(提升)和Gain(增益)
  - **Lift** =[TP/(TP+FP)] / [(TP+FN)/(TP+FP+FN+TN)]



纵坐标是lift，横坐标是正例集百分比。

真实情况	预测结果	
	正例	反例
正例	TP	FN
反例	FP	TN

- Lift(提升)和Gain(增益)
  - **Gain=** TP/(TP+FP)



纵坐标Gain，横坐标是正例集百分比。