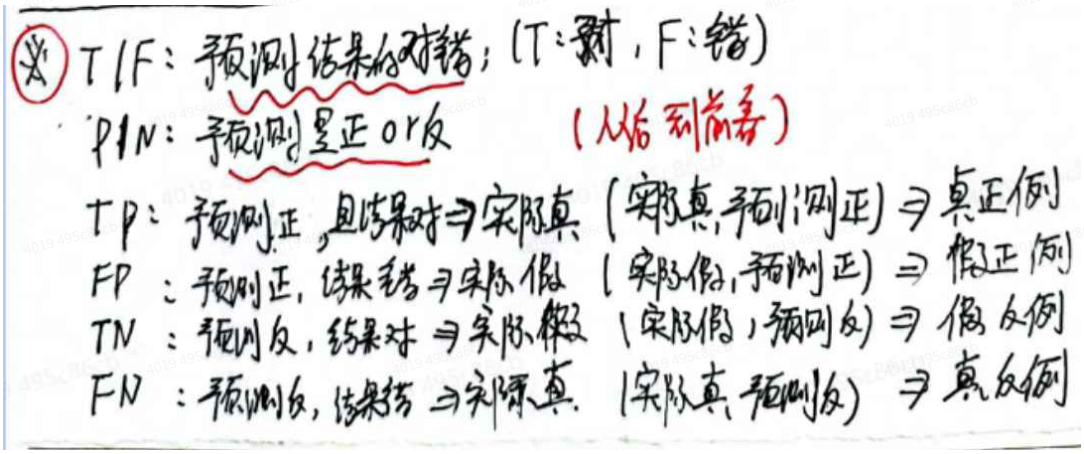
8.模型评估指标

1.混淆矩阵

混淆矩阵		真实值	
		Positive	Negative
预测值	Positive	TP	FP Type I
	Negative	FN Type II	TN



2.二级指标 ACC，PPV，TPR，TNR

2.1 准确率ACC：预测正确 / 所有

	公式	意义
准确率 ACC	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	分类模型所有判断正确的结果占总观测值的比重

预测正确：T = （TP + TN）

所有： （TP，TN，FP，FN）

2.2 精确率、查准率：Precision： （实际真，预测正）/ (预测正)

精确率 PPV	$\text{Precision} = \frac{TP}{TP + FP}$	在模型预测是Positive的所有结果中，模型预测对的比重

2.3 灵敏度、召回率、查全率：（实际真，预测正）/(实际真)

灵敏度 TPR	$\text{Sensitivity} = \text{Recall} = \frac{TP}{TP + FN}$	在真实值是Positive的所有结果中，模型预测对的比重

2.4特异性：（实际假，预测假）/ (实际假)

特异度 TNR	$\text{Specificity} = \frac{TN}{TN + FP}$	在真实值是Negative的所有结果中，模型预测对的比重

3.三级指标

3.1F1值

$$\frac{2}{F1} = \frac{1}{Recall} + \frac{1}{Precision}$$

4.ROC，AUC

4.1AUC是什么

在统计和机器学习中，常常用AUC来评估二分类模型的性能。AUC的全称是 area under the curve，即曲线下的面积。 通常这里的曲线指的是受试者操作曲线(Receiver operating characteristic, ROC)。**相比于准确率、召回率、F1值等依赖于判决阈值的评估指标，AUC则没有这个问题。**

对于二分类问题，预测模型会对每一个样本预测一个得分s或者一个概率p。 然后，可以选取一个阈值t，让得分s>t的样本预测为正，而得分s<t的样本预测为负。

这样一来，根据预测的结果和实际的标签可以把样本分为4类：

	正样本	负样本
预测为正	TP(真正例)	FP(假正例)
预测为负	FN(假负例)	TN(真负例)

随着阈值t选取的不同，这四类样本的比例各不相同。定义真正例率TPR和假正例率FPR为：

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$
$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

对于真正例率TPR，分子是得分>t （得分大于t就是预测为正）里面**正样本**的数目，分母是总的**正样本**数目。

而对于假正例率FPR，分子是得分>t里面**负样本**的数目，分母是总的**负样本**数目。

随着阈值t的变化，TPR和FPR在坐标图上形成一条曲线，这条曲线就是ROC曲线。

4.2AUC的概率解释

AUC常常被用来作为模型排序好坏的指标，**原因在于AUC可以看做随机从正负样本中选取一对正负样本，其中正样本的得分大于负样本的概率！**

4.2.1排序特性

根据上述概率解释，AUC实际上在说一个模型**把正样本排在负样本前面的概率**！

4.2.2对样本比例不敏感

在训练模型的时候，如果正负比例差异比较大，例如正负比例为1:1000，训练模型的时候通常要对负样本进行下采样。

当一个模型训练完了之后，用负样本下采样后的测试集计算出来的AUC和未采样的测试集计算的AUC基本一致，**因为采样均匀的，即 >s+的负样本和 <s+ 的负样本留下的概率是相同的，负样本的得分比例不会发生改变**，因此AUC值（**正样本的得分大于负样本的概率**）也不会发生改变！

相比于其他评估指标，例如准确率、召回率和F1值，负样本下采样相当于只将一部分真实的负例排除掉了，**然而模型并不能准确地识别出这些负例，所以用下采样后的样本来评估会高估准确率**；因为采样只对负样本采样，正样本都在，所以采样对召回率并没什么影响。这两者结合起来，最终导致高估F1值！