机器学习实验评分和分析教程-财务重点

5.任务3：实验评分和分析概念

正如我们在《自动机器学习入门教程概念》中了解到的一旦生成了模型，就必须对其性能进行评估。这些指标用于评估所构建模型的质量，以及应使用哪种模型得分阈值进行预测。有多种指标可用于评估二进制分类机器学习模型，例如接收器工作特性或ROC曲线，精度和召回率或Prec-Recall，Lift，Gain和KS图表仅举几例。每个指标评估机器学习模型的不同方面。以下概念适用于H2O无人驾驶AI中用于评估其生成的分类模型的性能的指标。在较高的层次上介绍了这些概念，为了在此处更深入地了解此处涵盖的每个指标，我们在此任务结束时还添加了其他资源。

二进制分类器

让我们看一下二进制分类模型。二元分类模型预测给定集合的元素属于哪两个类别（类）。在我们的示例中，这两个类别（类）是您住房贷款的**违约**，而**不是违约**。生成的模型应该能够预测每个客户属于哪个类别。

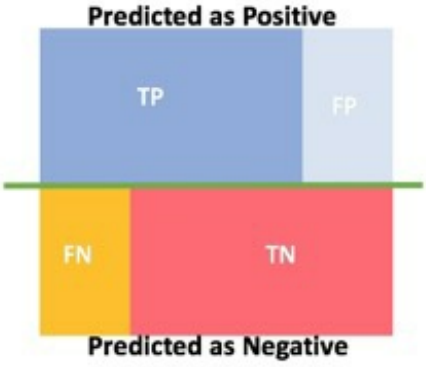
但是，还需要考虑其他两种可能的结果：误报和误报。在这些情况下，模型预测某人没有拖欠其银行贷款，但确实这样做。另一种情况是该模型预测某人拖欠其抵押贷款，但实际上却没有。总结果通过混淆矩阵可视化，混淆矩阵如下所示：

二进制分类产生四个结果：

预测为正：
真正= TP
假正= FP

预测为负：
真负= TN
假负= FN

混淆矩阵：



从这个混淆表中，我们可以测量错误率，准确性，特异性，敏感性和精确度，以及所有有用的指标，以测试我们的模型在分类或预测方面的表现。这些指标将在下一部分中定义和解释。

有趣的是，您可能想知道为什么命名为“ Confusion Matrix”？有人可能会说这是因为混淆矩阵可能非常令人困惑。除了笑话，混淆矩阵也称为**误差矩阵，**因为它使可视化模型的分类率（包括误差率）变得容易。心理学

中也使用术语“混淆矩阵”，牛津词典将其定义为“表示相对频率的矩阵，在该相对频率下，**许多刺激中的每一个都被误认为是其他刺激。**由某人在需要识别或识别刺激的任务中完成的。通过对这些数据的分析，研究人员可以提取因子（2），该因子指示受访者感知中相似性的潜在维度。例如，在颜色识别任务中，红色与绿色的相对频繁的**混淆**倾向于暗示道尔顿主义。” [1]换句话说，执行分类任务的人员多久将一项混淆为另一项。在ML的情况下，机器学习模型正在实施分类并评估模型混淆一个标签而不是人的标签的频率。

		True/Actual Class	
Predicted Class		Positive (P)	Negative (N)
	True (T)	TP (# of TPs)	FP (# of FPs)
	False (F)	FN (# of FNs)	TN (# of TNs)

鹏

解决分类问题的必要工具是ROC曲线或接收器工作特性曲线。ROC曲线直观地显示了二进制分类器的性能。换句话说，它“表明一个模型能够区分类别的程度” [2]和相应的阈值。继续以房地美（Freddie Mac）示例为例，输出变量或标签是客户是否将拖欠其贷款以及达到多少阈值。

一旦使用训练数据集构建并训练了模型，就可以通过分类方法（逻辑回归，朴素贝叶斯分类器，支持向量机，决策树，随机森林等）传递模型，从而得出每个客户都有违约。

ROC曲线针对每个可能的分类阈值绘制了灵敏度或真实阳性率（y轴）与1-Specific或错误阳性率（x轴）的关系图。分类阈值或决策阈值是模型将用来确定类别所属位置的概率值。该阈值充当类之间的边界，以从另一个类中确定一个类。由于我们要处理的是0到1之间的值的概率，因此阈值的示例可以是0.5。这告诉模型，低于0.5的任何事物都属于一个类，高于0.5的任何事物都属于另一类。可以选择阈值以使真实肯定最大化，同时使假阳性最小化。阈值取决于要应用ROC曲线的情况以及我们希望最大化的输出类型。

给定我们的预测贷款用例的示例，以下内容对混淆矩阵中的值进行了描述：

- TP = 1 =预测匹配结果某人确实拖欠贷款
- TN = 0 =预测匹配结果是某人未拖欠贷款
- FP = 1 =预测某人将违约，但实际上他们并未违约
- FN = 0 =预测某人未拖欠其银行贷款，但实际上已拖欠。

什么是敏感性和特异性？真实阳性率是真实阳性预测数除以所有阳性预测的比率。此比率也称为**召回率**或**灵敏度**，它的范围是0.0到1.0，其中0是最差的，1.0是最佳灵敏度。敏感度是模型对肯定情况的预测程度的度量。

真实否定率是真实否定预测数除以所有肯定预测的比率。该比率也称为**特异性**，范围是0.0到1.0，其中0是最差的，1.0是最佳特异性。特异性是衡量模型对否定情况正确预测的程度。它多长时间能正确预测一个否定情况。

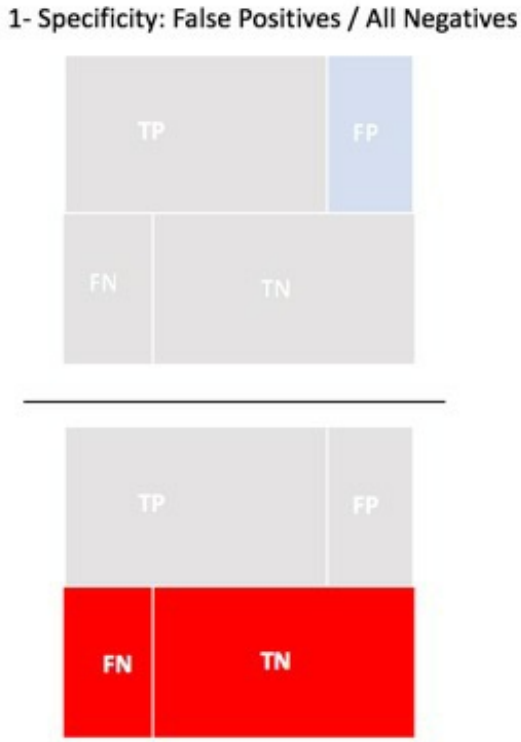
假阴性率是*1-特异性*，或者是假阳性率除以所有阴性预测的比率[3]。

下图说明了灵敏度，特异性和假阴性率的比率。

召回率 = **灵敏度** =真正率= TP /（TP + FN）

特异性 =真阴性率= TN /（FP + TN）

1-特异性 =假阳性率= 1-真阴性率= FP /（FP + TN）

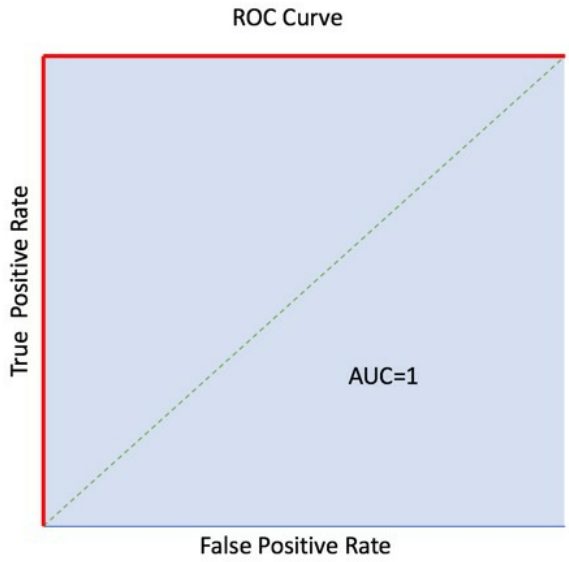


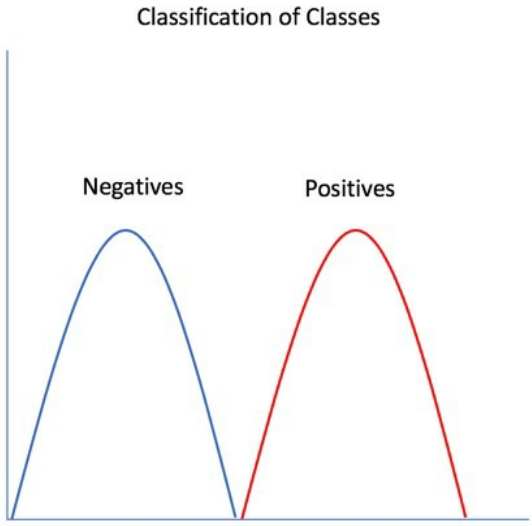
ROC曲线还可以通过量化其性能来告诉您模型的性能。得分由ROC曲线下的面积百分比确定，也称为曲线下面积或AUC。

以下是四种带有AUC的ROC曲线：

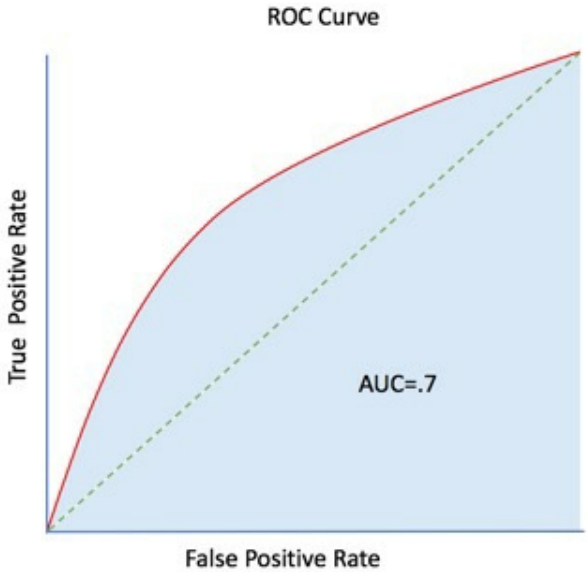
注意： ROC曲线越靠近左侧（AUC百分比越大），则模型在类之间进行分离的效果越好。

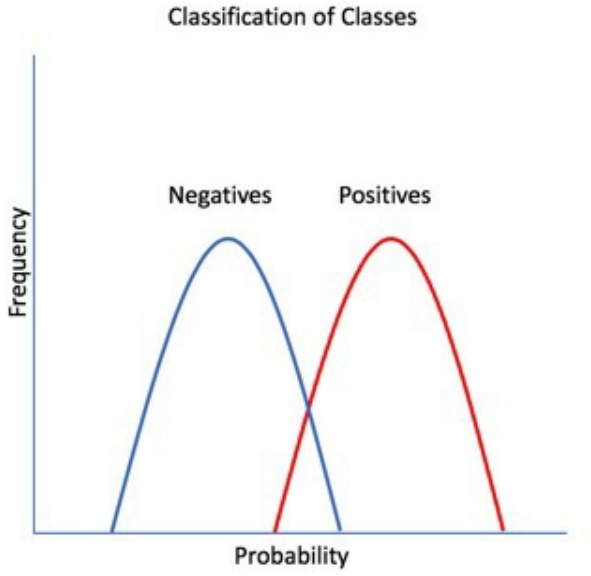
下方的完美ROC曲线（红色）可以100%的准确度分隔各个类，并且AUC为1.0（蓝色）：





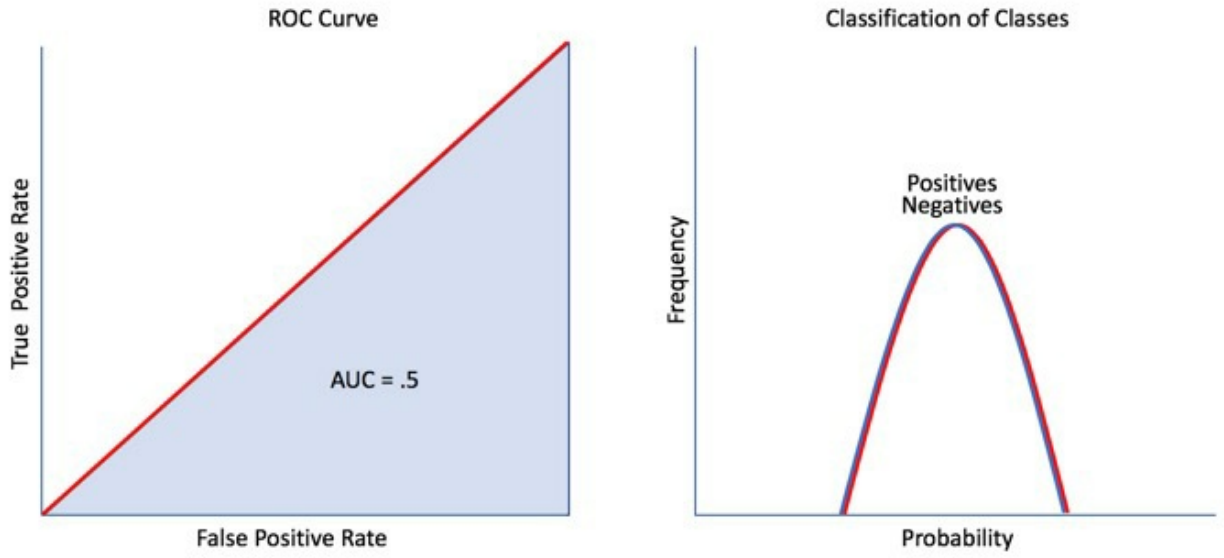
下方的ROC曲线非常靠近左角，因此它在分离AUC为0.7或70%的类别时表现出色：





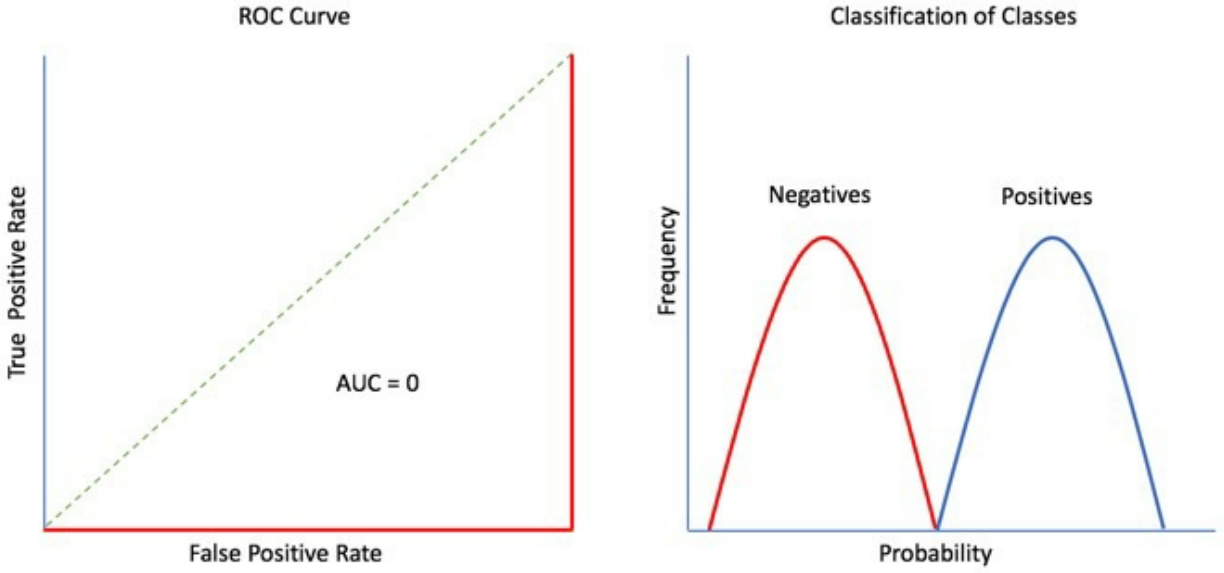
在超过70%的情况下，模型可以正确预测阳性和阴性结果，而在30%的情况下，模型会混合使用FP或FN。

该ROC曲线位于将图形分成两半的对角线上。由于它离左角较远，因此在区分类别时做得很差，这是最坏的情况，它的AUC为.05或50%：



AUC为0.5，表明我们的模型与随机模型一样好，后者有50%的机会预测结果。我们的模型并不比掷硬币好，模型可以正确预测结果的时间为50%。

最后，下面的ROC曲线代表了另一个完美的场景！当ROC曲线低于50%模型或随机机会模型时，则需要仔细检查模型。这样做的原因是，可能存在对负值和正值的错误贴标签，导致值反转，因此ROC曲线低于随机机会模型。尽管此ROC曲线在翻转时看起来具有0.0或0%的AUC，但我们却获得了1或100%的AUC。



ROC曲线是一个有用的工具，因为它仅关注模型能够区分类别的程度。“ AUC可以帮助表示分类器将随机选择的阳性观察结果的排名高于随机选择的阴性观察结果的概率” [4]。但是，对于很少发生预测的模型，较高的AUC可能会误导该模型正在正确预测结果。这就是精确度和召回率概念变得重要的地方。

精确召回

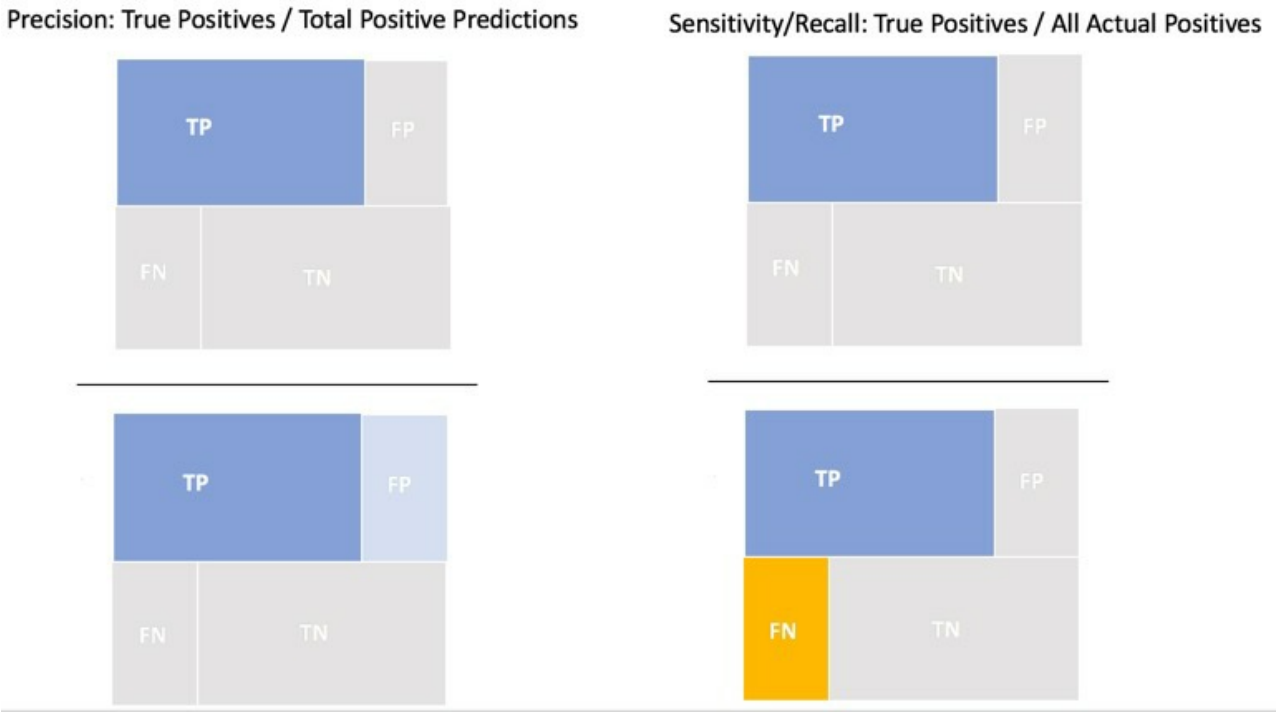
Precision-Recall曲线或Prec-Recall或**PR**是另一个评估从混淆矩阵得出的分类模型的工具。Prec-Recall是ROC曲线的补充工具，尤其是在数据集具有明显偏斜的情况下。Prec-Recall曲线绘制了每种可能的分类阈值的精度或正预测值（y轴）与灵敏度或真实正值（x轴）的关系。在较高的层次上，我们可以将精度视为对结果的准确性或质量的度量，而对召回率则是对模型所获得结果的完整性或数量的度量。Prec-Recall测量模型获得的结果的相关性。

精度是正确的肯定预测除以肯定预测的总数之比。该比率也称为**正预测值**，范围是0.0到1.0，其中0.0是最差的值，而1.0是最佳精度。精度更多地关注积极类别而不是消极类别，它实际上衡量的是正确检测正值（TP和FP）的可能性。

精度 =真实的阳性预测/阳性预测的总数= TP /（TP + FP）

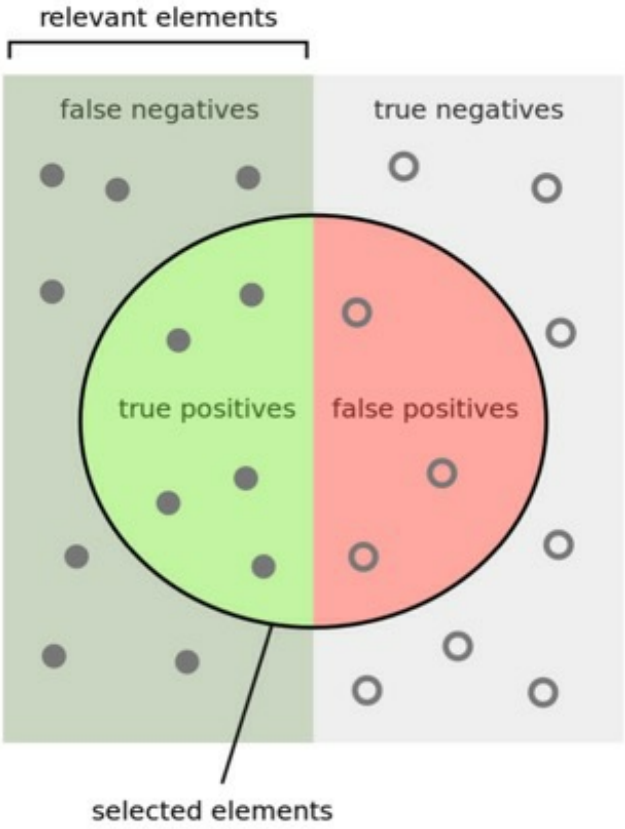
如ROC部分所述，**召回**率是真实阳性率，是真实阳性预测数除以所有阳性真实预测之比。召回率是实际肯定预测的度量。它告诉我们在模型测试期间，所有可用的阳性样本中产生了多少正确的阳性结果。

召回率 = **灵敏度** =真正率= TP /（TP + FN）

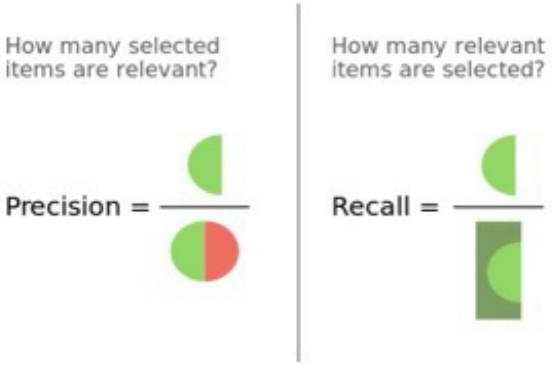


下面是可视化Precision和Recall的另一种方式，该图像是从https://commons.wikimedia.org/wiki/File:Precisionrecall.svg借用的。

通过非线性插值[5]连接所有精确召回点，即可创建精确召回曲线。召回前图分为两个部分："好"和"差"。可以在该图的右上角找到"良好"性能，在左下角找到"较差"性能，请参见下图查看完美的重调用前图。该划分由基线生成。Prec-Recall的基准由正数（P）与负数（N）之比确定，其中y = P /（P + N），该函数表示具有随机性能水平的分类器[6]。当数据集平衡时，基线值为y = 0.5。如果数据集不平衡，其中P的数量大于N的数量，则将相应地调整基线，反之亦然。



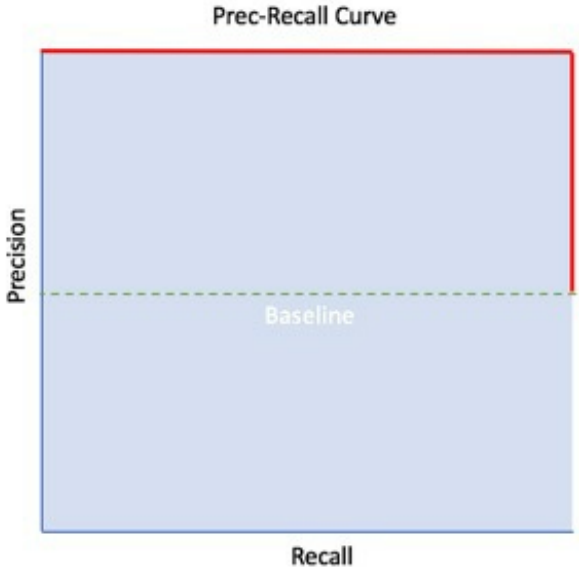
完美的精确调用曲线是两条直线（红色）的组合。该图告诉我们该模型没有预测错误！换句话说，假设基线为0.5，则没有误报（完美的精度）和没有误报（完美的回忆）。



与ROC曲线类似，我们可以使用曲线下的面积或AUC来帮助我们比较模型与其他模型的性能。

注意：" Prec-Recall曲线"越靠近右上角（AUC百分比越大），模型就越能正确地预测真实阳性。

下面的红色Prec-Recall曲线的AUC约为0.7（蓝色），相对基线为0.5：



最后，此Prec-Recall曲线表示最坏的情况，其中该模型生成100%的假阳性和假阴性。此Prec-Recall曲线的AUC为0.0或0%：

从Prec-Recall图中可以得出一些有助于评估模型性能的指标，例如准确性和F_β得分。这些指标将在概念的下一部分中更深入地说明。只需注意，准确度或ACC是正确预测的比率除以预测总数，而F_β是召回率和精确度的调和平均值。

当以Pre-Recall精度查看ACC时，必须注意以下积极观察，即ACC不能执行均衡性很好的数据集。这就是为什么**F分数**可用于解释Prec-Recall中偏斜的数据集的原因。

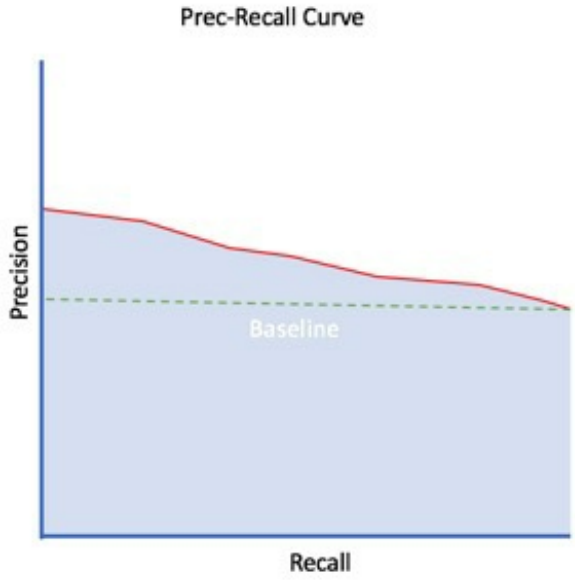
当您考虑模型对于阳性案例的准确性时，您需要了解以下几点：

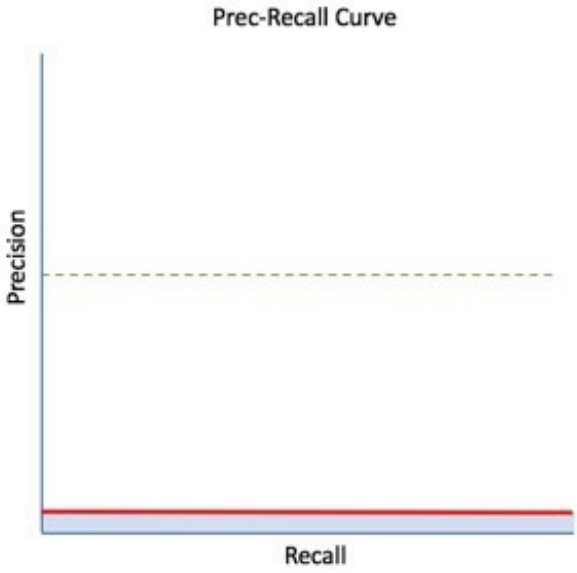
- 多久正确一次？
- 什么时候错了？为什么？
- 是因为您的误报过多吗？（精确）
- 还是因为您的假阴性过多？（召回）

还可以考虑各种F1分数，即F1，F2和F0.5。1、2和0.5是赋予查全率和准确性的权重。例如，F1表示精度和查全率具有相等的权重，而F2给出的查全率比精度高，而F0.5给出的查准率比召回率高。

Prec-Recall是考虑分类器的好工具，因为它是类分布中较大偏差的绝佳选择。使用精确度和查全率专注于小的阳性分类—当阳性分类较小且能够正确检测阳性样本的能力是我们的主要重点（正确检测阴性样本对问题的重要性不大）时，我们应该使用精确度和查全率。

如果您使用的是准确度模型度量标准，而Prec-Recall出现问题，那么您可以考虑使用对数损失模型度量标准。





GINI，ACC，F1 F0.5，F2，MCC和对数丢失

ROC和Pre-Recall曲线对于测试二进制分类器非常有用，因为它们提供了每个可能分类阈值的可视化。从这些图中，我们可以得出单个模型指标，例如ACC，F1，F0.5，F2和MCC。还有其他单个度量标准可以同时用于评估模型，例如GINI和Log Loss。以下将讨论模型得分ACC，F1，F0.5，F2，MCC，GINI和对数损失。模型得分是ML模型最优化的。

基尼

基尼系数是一种建立良好的方法，可以量化频率分布值之间的不等式，并且可以用来衡量二进制分类器的质量。基尼系数为零表示完全相等（或完全没用的分类器），而基尼系数为1表示最大不平等（或完全分类器）。

基尼系数基于洛伦兹曲线。洛伦兹曲线将真实的阳性率（y轴）绘制为总体百分位数（x轴）的函数。

洛伦兹曲线代表由分类器代表的模型的集合。曲线上的位置由特定模型的概率阈值给出。（即，较低的分类概率阈值通常会导致更多的真实阳性，但也会导致更多的假阳性。）[12]

基尼指数本身与模型无关，仅取决于从分类器获得的分数（或概率）的分布确定的洛伦兹曲线。

准确性

精度或ACC（不要与AUC或曲线下的面积相混淆）是二进制分类问题中的单个度量。ACC是正确预测的比率除以预测总数。换句话说，模型可以正确识别真假和真假的程度如何。精度在0到1的范围内测量，其中1是完全精度或完全分类，而0是精度差或分类[8]。

使用混淆矩阵表，可以按以下方式计算ACC：

精度 =（TP + TN）/（TP + TN + FP + FN）

F分数：F1，F0.5和F2

F1分数是分类准确性的另一种度量。它代表精度和查全率的谐波平均值。F1在0到1的范围内测量，其中0表示没有真实的正数，而1既没有假阴性也没有假正数或没有完美的精度和召回率[1]。

使用混淆矩阵表，可以按以下方式计算F1分数：

F1 = 2TP /（2TP + FN + FP）

F05公式：
F0.5 = 1.25（（精度）（调用）/ 0.25精度+调用）

其中：
精度是正观测值（真实正值），模型会从其标记为正值的所有观测值中正确识别（真实正值+虚假正值）。回忆是从所有实际阳性病例（真实阳性+假阴性）中正确识别的阳性观察结果（真实阳性）[15]。

该**F2得分**是的精度和召回（给定的阈值）的加权调和平均值。与F1得分不同，F1得分对准确性和回忆性的重视程度也不同，F2得分对回忆性的重视程度大于对准确性的重视程度。对于误报被认为比误报更糟糕的情况，应该给予更大的重视。例如，如果您的用例是预测哪些客户会流失，则您可能认为误报比误报更糟。在这种情况下，您希望您的预测能够捕获所有会流失的客户。这些客户中的一些可能没有搅动的风险，但是他们得到的额外关注无害。更重要的是，没有任何客户有可能遭受搅动的风险[15]。

我的客户中心

MCC或Matthews相关系数，用于衡量二进制分类的质量[1]。MCC是观察到的和预测的二进制分类之间的相关系数。MCC的测量范围是-1和+1之间，其中+1是理想的预测，0优于随机预测，而-1是所有不正确的预测[9]。

使用混淆矩阵表，可以按以下方式计算MCC：

MCC =（TP * TN- FP * FN）/ [（TP + FP）*（FN + TN）*（FP + TN）*（TP + FN）] ^½

对数损失（对数损失）

对数损失度量可用于评估二项式或多项式分类器的性能。与AUC着眼于模型对二进制目标的分类效果不同，logloss评估模型的预测值（未校准的概率估计）与实际目标值的接近程度。例如，模型是倾向于为正类别指定较高的预测值（如.80），还是显示识别正类别并赋予较低的预测值（如.50）的能力较弱？对数损失为0的模型将是理想的分类器。当模型无法做出正确的预测时，对数损失会增加，从而使模型成为不良模型[11]。

二元分类方程：

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N} w_i(y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i))$$

多类分类方程：

$$Logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} w_i(y_{i,j} \ln(p_{i,j}))$$

哪里：

- N是相应数据框的总行数（观测值）。
- w是每行用户定义的权重（默认为1）。
- C是类的总数（对于二进制分类，C = 2）。
- p是分配给给定行（观察值）的预测值（未校准概率）。
- y是实际目标值。

无人驾驶AI诊断会计算ACC，F1，MCC值，并在每个ROC和召回前曲线中绘制这些值，从而更容易为生成的模型识别最佳阈值。此外，它还可以计算模型的对数损失得分，从而使您可以快速评估生成的模型是否是好的模型。

让我们回到评估模型的指标结果。

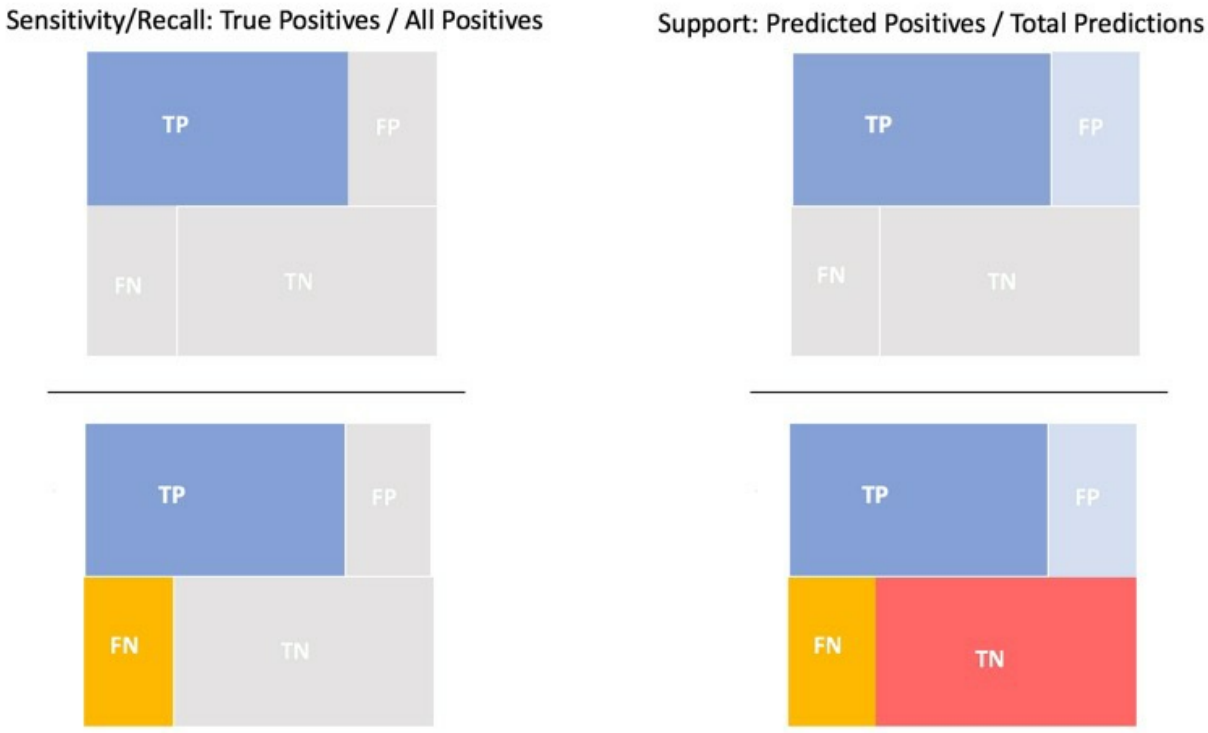
增益和提升图

增益图和提升图通过查看训练模型与随机模型（或无模型）获得的结果之间的比率来衡量分类模型的有效性[7]。增益和提升图可帮助我们评估分类器的性能，并回答一些问题，例如所捕获的数据集中有多少百分比具有正响应，取决于所选样本百分比。此外，我们可以探索与随机模型（或没有模型）相比，使用模型可以做的更好[7]。

我们可以认为获得收益的一种方式是“对于预测结果的每一步，不确定性的水平都会降低。不确定性的下降是导致知识获得的熵的损失” [15]。增益图表绘制了真实阳性率（敏感性）与预测阳性率（**支持**）的关系，其中：

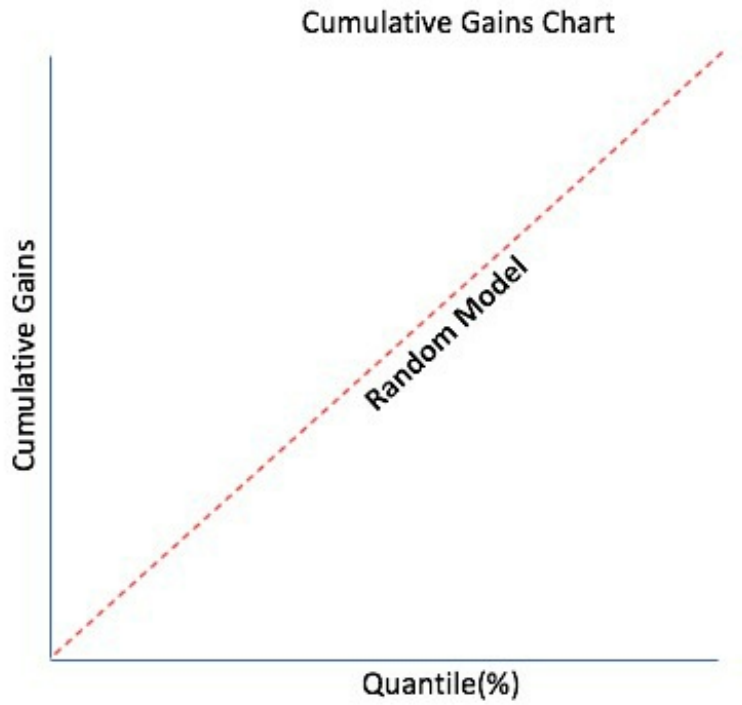
灵敏度 = **召回**率=真阳性率= TP /（TP + FN）

支持 = **预测阳性率** = TP + FP /（TP + FP + FN + TN）



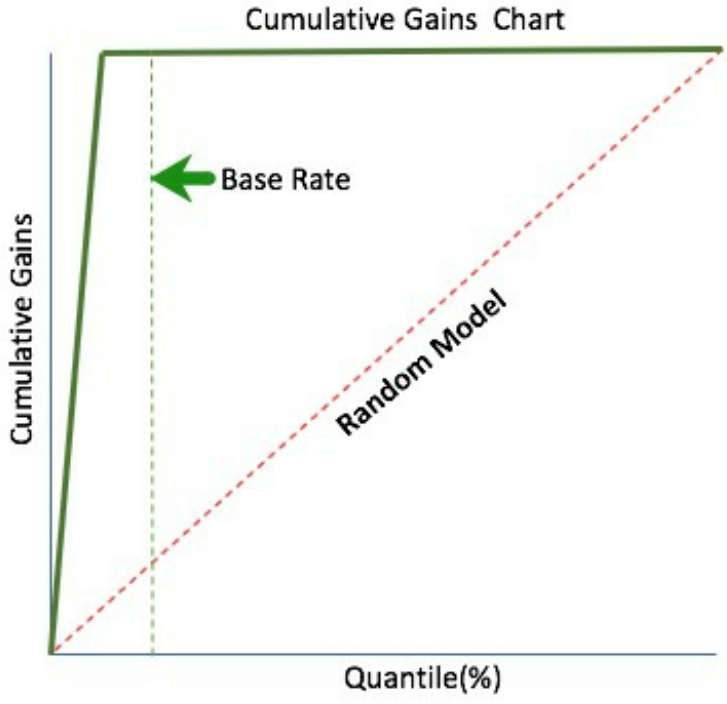
为了更好地形象化阳性反应与选定百分比样本相比的百分比，我们使用**累积增益**和**分位数**。通过采用预测模型并将其应用于作为原始数据集的子集的测试数据集，可以获得累积收益。预测模型将以概率为每个案例评分。然后按预测分数对分数进行升序排列。分位数取案件总数（有限数量），并将有限集划分为几乎相等大小的子集。百分位数是从第0个百分位数到第100个百分位数绘制的。然后，我们绘制直到每个分位数的累计病例数，从出现概率最高的0%的阳性案例开始，直到得分最低的阳性案例的100%为止。

在累积收益图表中，x轴显示了测试数据集中病例在病例总数中所占的百分比，而y轴显示了以分位数表示的阳性反应的百分比。如前所述，由于概率是按升序排序的，因此我们可以查看在10%或20%中发现的预测阳性病例的百分比，以此来缩小我们感兴趣的阳性病例的数量。可以将预测模型的模型与随机模型（或没有模型）的模型进行比较。随机模型在下面以红色表示，这是随机抽样的最坏情况。



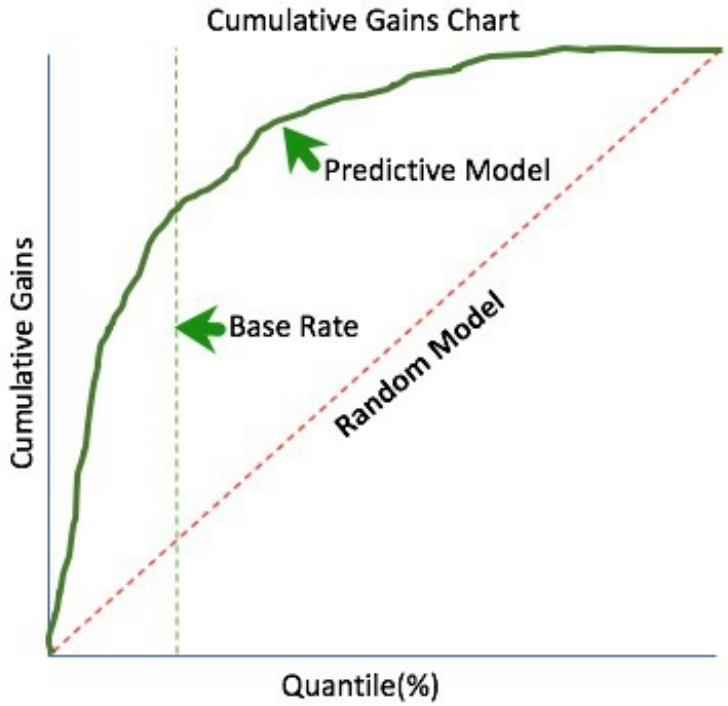
我们如何确定与随机模型有关的最佳情况？为此，我们需要先确定基本费率。基本费率设置最佳曲线的极限。最佳收益始终由基本利率控制。下图（绿色虚线）显示了基本费率的示例。

- **基本费率**定义为：
- **基本速率** =（TP + FN）/样本大小



上面的图表代表了假设基准利率为20%的累积收益图表的最佳情况。在这种情况下，在达到基本比率之前，所有阳性病例都已被识别。

下图显示了预测模型的示例（绿色实线）。我们可以看到与随机模型（红色虚线）相比，预测模型的性能如何。现在，我们可以选择一个分位数，并确定相对于整个测试数据集，四分位数以上的阳性案例所占的百分比。

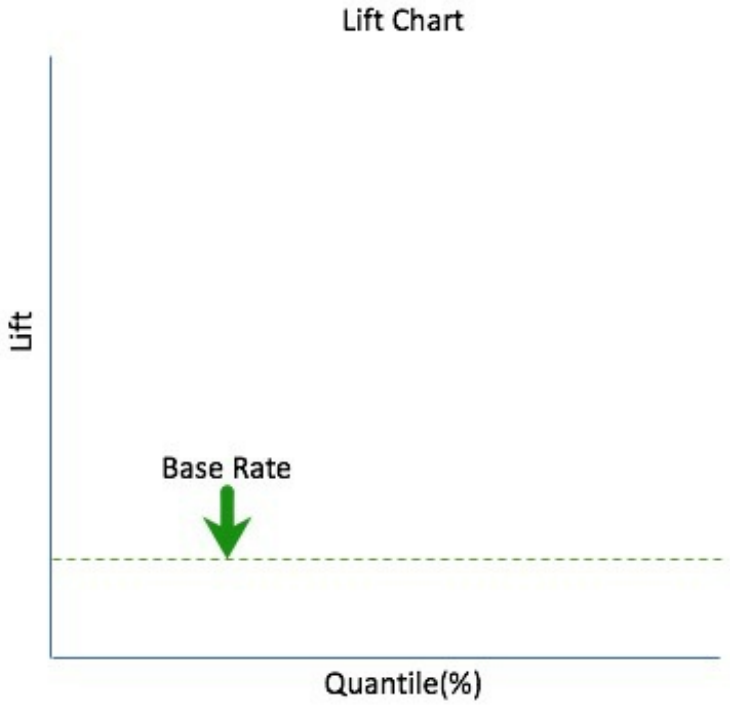


提升可以帮助我们回答以下问题：与随机模型（或没有模型）相比，预测模型可以做的更好。提升度是对预测模型有效性的度量，该预测模型的计算方式是使用一个模型和一个随机模型（或没有一个模型）获得的结果之间的比率。换句话说，在给定的分位数上，增益%与随机期望%的比率。第x个分位数的随机期望为x%[16]。

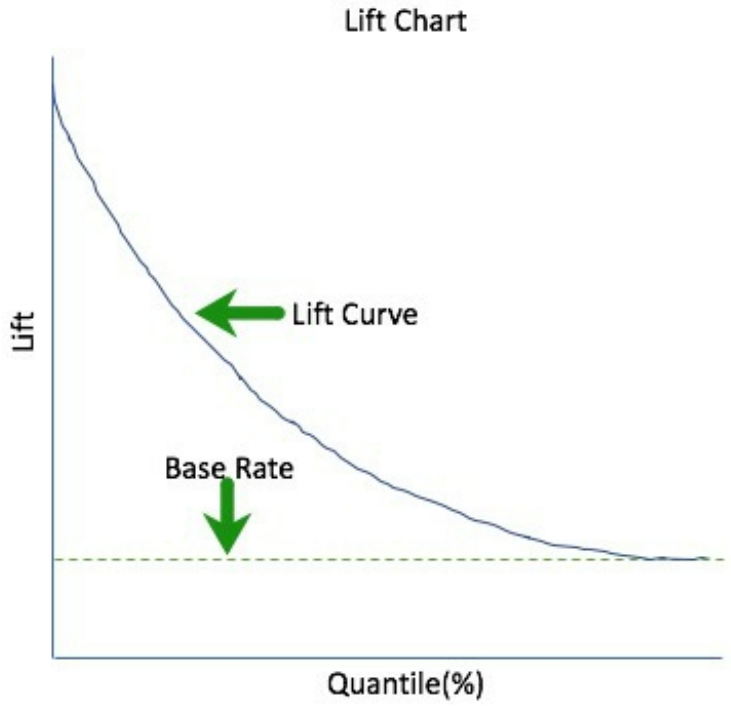
提升 =预测率/实际率

在绘制升力时，我们还将其相对于分位数进行绘制，以帮助我们直观地看出，由于升力图是从累积收益图中得出的，因此有可能发生积极的情况。通过确定我们的模型预测的结果与使用随机模型（或没有模型）的结果之间的比率来计算升力曲线的点。例如，假设随机模型的基本利率（或假设阈值）为20%，我们将获得20%分位数的累积增益百分比X并除以20。我们对所有分位数都这样做，直到得到完整的升力曲线。

我们可以从基本费率开始如下所示的升程图，回想一下基本费率是目标阈值。



当查看最高分位数的累积提升X时，这意味着当我们根据模式从总测试用例中选择分位数说20%时，我们可以期望X / 20乘以总数通过从随机模型中随机选择20%发现的阳性病例数。

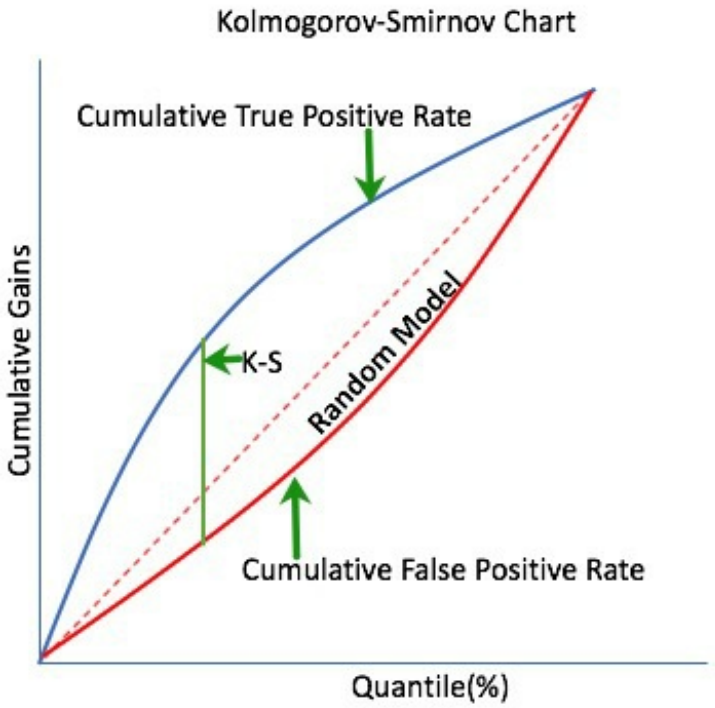


KS图

Kolmogorov- Smirnov或KS通过测量阳性或阴性之间的分离程度来验证或测试数据，从而测量分类模型的性能[13]。“如果分数将人群分为两个独立的组，则KS为100，其中一组包含所有正值，另一组包含所有负值。另一方面，如果模型无法区分正值和负值，则好像“模型会从总体中

随机选择病例。KS将为0。在大多数分类模型中，KS会落在0到100之间，值越大，模型将阳性病例与阴性病例分开的效果就越好。” [14]。

KS统计量是响应者的累积百分比或1（累积的真实阳性率）与无响应者的累积百分比或0（累积的假阳性率）之间的最大差。KS统计的重要性在于，它有助于理解应该针对哪些人群以获得最高响应率（1）[17]。



参考文献

[1] 混淆矩阵定义“心理学词典”

[2] 走向数据科学-了解AUC-ROC曲线

[3] ROC简介

[4]解释ROC曲线和曲线下曲线（AUC）

[5] 精确召回简介

[6] Tharwat，应用计算和信息学（2018年）

[7] 模型评估分类

[8] Wiki准确性

[9] Wiki F1分数

[10] Wiki马修的相关系数

[11] Wiki日志丢失

[12] H2O的GINI指数

[13] H2O的Kolmogorov-Smirnov

[14] 模型评估-分类

[15] 什么是机器学习中的信息获取

[16] 提升分析数据科学家的秘密武器

[17] 机器学习评估指标分类模型

更深入的潜水和资源