**打卡一**

**赛题背景：**

赛题以金融风控中的个人信贷为背景，要求选手根据贷款申请人的数据信息预测其是否有违约的可能，以此判断是否通过此项贷款，这是一个典型的分类问题。

**赛题数据：**

赛题以预测用户贷款是否违约为任务，数据集报名后可见并可下载，该数据来自某信贷平台的贷款记录，总数据量超过120w，包含47列变量信息，其中15列为匿名变量。为了保证比赛的公平性，将会从中抽取80万条作为训练集，20万条作为测试集A，20万条作为测试集B，同时会对employmentTitle、purpose、postCode和title等信息进行脱敏。





**评判标准：**

提交结果为每个测试样本是1的概率，也就是y为1的概率。评价方法为AUC评估模型效果（越大越好）。

AUC（Area Under Curve）被定义为[ROC曲线](https://baike.baidu.com/item/ROC%E6%9B%B2%E7%BA%BF)下与坐标轴围成的面积，显然这个面积的数值不会大于1。又由于ROC曲线一般都处于y=x这条直线的上方，所以AUC的取值范围在0.5和1之间。AUC越接近1.0，检测方法真实性越高;等于0.5时，则真实性最低，无应用价值。

其中，ROC曲线全称为受试者工作特征曲线 （receiver operating characteristic curve），它是根据一系列不同的二分类方式（分界值或决定阈），以真阳性率（敏感性）为纵坐标，假阳性率（1-特异性）为横坐标绘制的曲线。

ROC曲线的横坐标是伪阳性率（也叫假正类率，False Positive Rate），纵坐标是真阳性率（真正类率，True Positive Rate），相应的还有真阴性率（真负类率，True Negative Rate）和伪阴性率（假负类率，False Negative Rate）。这四类指标的计算方法如下：

　　（1）伪阳性率（FPR）：判定为正例却不是真正例的概率，即真负例中判为正例的概率

　　（2）真阳性率（TPR）：判定为正例也是真正例的概率，即真正例中判为正例的概率（也即正例召回率）

　　（3）伪阴性率（FNR）：判定为负例却不是真负例的概率，即真正例中判为负例的概率。

（4）真阴性率（TNR）：判定为负例也是真负例的概率，即真负例中判为负例的概率。