本题解题思路是：

观察到附件2相比附件1，第一张sheet表中缺少两列数据——是否违约以及信誉评级

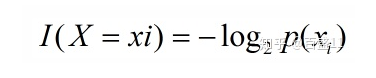
银行做出信贷策略肯定是需要使用到这两个信息，于是使用集成学习方法——决策树和随机森林来预测这两列数据！

随机森林的基本原理：

信息、熵和信息增益：

这三个概念是决策树使用特征来分类时确定特征选取顺序的依据

在机器学习中，若带分类的事物集合可以根据多重标准进行划分，则那个类可以定义为：



其中I(x)表示Xi发生的事件所包含的信息量，，P(xi)代表Xi发生的概率

于是接着引出了熵的概念以及定义：



熵越大，代表这个类别的不确定性越大，反之越小

而信息增益是机器学习中用来选择特征的指标，信息增益越大，代表这个特征选择性越好

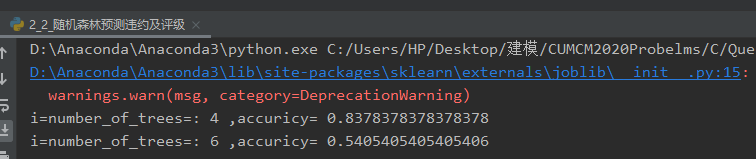


本题的训练集和验证集采用的是附件1中的123家企业的【是否违约】以及【信誉评级】和一些处理后的特征数据，其中训练集和验证集按照7：3随机抽样，唯一的验证结果是【是否违约】和【信誉评级】的指标，因为要保证每棵树的训练本要有交集，因此使用的是又放回的抽样方式并且设置了随机种子

这边采用的是“最大熵“的方法，经过训练集和验证集的测试，得出训练结果正确率最高的情况是：36棵子决策树以及最小样本叶片是4

然后使用sklearn模块里面的fit方法进行训练，使用predict进行预测，使用accuracy\_score来比对预测和验证集的正确程度

最后将模型保存到本地，使用随机森林的回归方法，预测同样数据处理后的附件2，最后将预测后的数据写到处理后的文件中



上图分别是【是否违约】和【信誉评级】的预测正确率结果