**（科技型）中小企业成长性评价模型**从六个维度出发对企业进行评价，包 括技术创新能力、盈利能力、发展潜力、管理能力、营运效率、企业信用风险。（六维）

这些是六维里面的，可能有这些属性，编程的时候可以先写着，到时候改

专利（%）

博士研究生占比（%）

企业财务情况（%）

三月内盈利状况（%）

固定资产情况（%）

其它方面（%）

百分比通过后面建模计算



代码先写六维，下面的属性先不管

建模方法介绍

**第一部分：基础模型**。成长性评价模型的基础为 **GBRT 算法**，其它两大模 块在此基础上建立。GBRT 算法具有精度高，泛化能力强，能处理非线性数据，能处理多特征类型，对数据的清洁程度要求相对较低等优点，非常适合成长性评价模型的需求和数据环境。



**第二部分：集成学习模块**。集成学习模块采用 stacking 算法，进一步提 升模型的精度，稳定性及泛化能力。在有标记样本的数量较少的情况下，可能同时存在多个模型在训练集上具有良好的表现，并且表现相对较差的模型在实际使用时，可能相对训练集上表现更好的模型具有更高的精度。同时，样本量较小的情况下，模型也更容易受极端值，异常点等影响。通过集成学习，综合多个模型的预测结果，让模型的预测能力更为稳定可靠。

**第三部分：半监督学习模块**。半监督学习模块采用 Tri-training 算法，充分利用未标记样本的信息。传统建模方法训练模型不使用未标记样本，但实际上，未标记样本中同样存在大量信息可用于训练模型，半监督学习可以充分利用这些信息，进一步保障模型性能。

以 GBRT 为基模型，在数据规模和质量提升后，可以采用多种技术手段进一步提升模型的预测能力和稳定性。可以从以下两个角度出发：

半监督学习——半监督学习可以充分利用未标记样本信息，提升模型预测能力。如协同训练(Co-train)，其基本原理是首先基于训练集产生两个不同的模型（如 GBRT 和神经网络），两个模型同时对测试集进行预测，比较预测结果，挑选预测结果最可信的样本，将预测结果作为该样本的标签，添加进训练集，根据扩大后的训练集训练出新的模型，然后重复此过程。集成学习——集成学习可以通过集成多个模型提升模型性能。如 stacking模型，其基本原理是基于训练集训练出多个模型，将每个模型的输出作为输入，训练出一个新的模型，新模型的输出作为模型整体的输出。通过这一过程可以让多个模型互相修正，使模型的预测结果更稳定可靠。

在模型构建过程中，采用 10 折交叉验证检验模型预测性能，即每次抽取 十分之九的样本进行建模，对余下的十分之一的样本进行预测，观察预测效果，重复十次。验证结果如下，mae(mean absolute error)代表平均绝对误差，test-mae-mean 与 test-mae-std 分别代表十次交叉验证获得的 mae 的均 值与标准差。