DOI:10.13768/j.cnki.cn11-3793/f.2023.0933

基于GBRT算法的中小企业成长性评价模型研究

余 屹[1] 艾孜尔江·艾尔斯兰[2] 戴兆君[1] 廖文君[1] 沈颂东[1] 梁子浚[2*]

（1.珠海科技学院金融与贸易学院，广东 珠海 519090；

2.澳门大学科技学院，澳门 519000）

摘 要：本文首先对我国中小企业的特征进行了分析，搭建了能够精准评价中小企业成长的指标体系，包含盈利能力、信用风险、营运效率、管理能力、发展潜力以及技术创新能力六个维度的指标。然后，利用指标体系建立了中小企业成长性评价模型，以研究各维度对企业成长的影响程度。最后，以全国中小企业股份转让系统中的近2千家企业为例，使用GBRT算法实证模拟验证了模型的有效性和实用性。

关键词：GBRT算法；成长性评价模型；投资决策

一、引言

中小企业在稳定社会、提高就业率以及促进市场发展等方面具有不容小觑的作用。成长性是指企业持续发展的能力，通过对企业成长性的评价，管理者可以及时地发现并解决企业存在的问题，提高企业管理水平和自我修正能力。此外，评价结果的好坏将直接影响投资者做出的投资决策。所以，对企业的成长性进行准确的评价，可以达到多方共赢的效果。

现有学者对中小企业成长性的评价主要从企业内外部因素进行探讨，这为本文研究提供了重要的参考价值。外部因素指政治、技术和市场竞争等环境影响。Astrakhan[1]等论证了政府通过利好政策和相关法律的支持，为企业打造出极佳的外部发展环境；成璐璐[2]等通过对市场竞争环境变化的分析，得出技术创新对企业的发展壮大有较大的影响，使企业能够在市场竞争中处于领先地位。而在内部因素上，学者们更关注企业的财务和融资等。孔镜翔[3]从中小板和创业板中筛选了近10年的企业作为样本，借助SPSS软件功能实现了对企业成长性的评价，通过实证分析，挖掘出企业中高管的学历对企业成长的影响程度。

基金项目：广东省教育厅广东省重点建设学科科研能力提升项目（项目号：2021ZDJS137）《粤港澳大湾区数字金融与产业链重构理论和实践研究》

在设计评价体系时，大多数现有研究忽略了企业成长系统的复杂性，即没有考虑到企业在成长时可能会受到的各类影响间的相互关系。基于此，本文提出全新的企业成长性模型框架，在此基础上引用GBRT算法，通过实证模拟，检验模型的精确性，为企业提供有价值的参考依据。

二、基于GBRT算法的中小企业评价模型研究

（一）评价模型

GBRT（Gradient Boost Regression Tree）算法是一种迭代的回归树算法，会将所有回归树的结论累加起来作为阶段性结果。最终结果会由迭代多棵树来共同决策，其核心是每一棵树都是学习之前所有树的结论和残差。

其中，回归树的整理流程基本如下示：

输入：训练数据集D；

输出：回归树f（x）；

在训练数据集所在的输入空间中，递归地将每个区域划分为两个子区域并决定每个子区域上的输出值，构建二叉决策树：

1.选择最优切分变量j与切分点s，求解

$$\min_{j,s}\left[\min_{e_1}\sum_{x_i\in R_1(j,s)}(y_i=c_1)^2+\min_{e_2}\sum_{x_i\in R_2(j,s)}(y_i=c_2)^2\right] \quad (1)$$

遍历变量j，对固定的切分变量j扫切分点s，选择使式（2）达到小值的对（j，s）。

2.用选定的对（j，s）划分区域并决定相应的输出值：

$$R_1(j,s)=\{x\mid x^{(j)}\le s\},R_2(j,s)=\{x\mid x^{(j)}>s\} \quad (2)$$

$$\widehat{C_m}=\frac{1}{N_m}\sum_{x_i\in R_m(j,s)}y_t\ ,x_i\in R_m\ ,m=1,2 \quad (3)$$

3.继续对两个子区域调用步骤（1），（2），直至满足停止条件，

4.将继续输入空间划分为M个区域R_1，R_2，…，R_M，生成决策树：

$$f(x)=\sum_{m=1}^{M}\widehat{C_m}I(x\in R_m) \quad (4)$$

当完成回归树流程生成对应的决策树后，使用loss函数的梯度近似残差，解决残差计算问题；然后，以

合残差的近似值利用线性搜索估计叶结点区域的值，使损失函数极小化，得到最终模型$\widehat{f(x)}$。

$$-\left[\frac{\partial L(y,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \quad (5)$$

$$\widehat{f(x)} = f_M(x) = \sum_{m=1}^{M}\sum_{j=1}^{J} C_{mj} I(x \in R_{mj}) \quad (6)$$

上述步骤即为中小企业成长性评价模型利用GBRT算法的原理，具有强大的预测能力，不仅可以处理不同类型的数据，对空间外的异常点处理效果也非常显著。

（二）评价方法

本文所研究的中小企业评价模型分三步。

首先，先将GBRT算法作为模型基础，其他两大模块在此基础上建立。GBRT算法是一种集成学习技术，它是多个决策树结合形成的预测模型，具有精度高、泛化能力强、处理非线性数据等特点，非常适合成长性评价模型的使用。

表1　GBRT模块

历史数据			GBRT →	待评估企业数据		
企业上报数据	公开数据	专家打分		企业上报数据	公开数据	成长价值分

其次，将处理后的数据交由集成学习模块使用stacking算法进一步处理，其基本原理是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作为整体的输出。这一过程能提升模型的精度、稳定性及泛化能力，让模型的预测能力更为稳定可靠。

最后，通过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，提升模型预测能力。如协同训练（Co-train），是基于训练集产生两个不同的模型（如GBRT和神经网络）同时对测试集进行预测，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。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。

表2　半监督学习模块

历史数据			+	待评估企业数据		Co-training →	待评估企业数据		
企业上报数据	公开数据	专家打分		企业上报数据	公开数据		企业上报数据	公开数据	成长价值分

此外，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。该验证标准差较小，预测性能稳定可靠。即便迭代次数较少，依旧可以实现GBRT算法性能迅速提升并趋于稳定的效果。因此，该模型以GBRT算法为基础，在数据规模和质量提升后，通过集成学习和半监督学习模块，能够进一步提升模型的预测能力和稳定性，具有研究意义。

（三）实证模拟

1.数据来源

本文选取全国中小企业股份转让系统中的近2千家中小企业作为分析对象。结合中小企业在系统上所核算的财务、管理、营运等数据，加之企业或相关政府部门所公示的该公司的信用风险、知识产权等信息的量化数据，形成导入模型的基本数据。

2.评价指标说明

在已有的研究基础上，充分考虑全国中小企业成长特点，在满足GBRT算法要求的前提下，分别从盈利能力、营运效率等六个维度遴选出中小企业成长过程的主要影响因素，科学合理地构建中小企业成长性评价指标体系，如表3所示。

3.实证结果分析

本次实验在获得原始数据后对数据进行清洗，并进行重新审查和校验，对重复信息、错误数据进行纠正，确保从系统中数据的一致性。处理共得1700条数据，每条数据代表一个企业，特征是评价模型框架对应的六维和企业对应的总分。

将处理后的数据导入模型，结合本文所述操作，通过GBRT等算法的递进使用对1700家中小企业成长性进行评价。实证分析用Python作为开发语言，通过

表3　成长性评价模型框架表

信用风险	营运效率	管理能力	技术创新能力	盈利能力	发展潜力
失信风险	流动比例	管理人员数量	核心技术来源	主营收入	市场占用
欠税信息	资产信息	年龄构成	产品特征	毛/净利润	竞争对手
抵押信息	负债信息	创业经历	技术实践阶段	项目数量	知识产权
经济诉讼	产品周期	学历分布	研发投入	资产净值	政策扶持
…	…	…	…	…	…

scikit-learn（机器学习和数据挖掘）、matplotlib（绘制各种静态、动态、交互式图表和图形）和numpy（科学计算和数值分析）实现数据处理和建模。

首先，GBRT算法的估计可以优化侧重于通过生长多个决策树来最小化损失函数，即优化包括在树的每个节点找到最小化损失函数的最优分割，并更新、分配给每个特征的权重。

其次，将基本数据所具有6个特征进行梯度增强模型拟合到训练数据的结果是近似目标变量和6个特征之间的关系的预测模型。也就是构建多个决策树，以目标变量的分段常数近似的方式拟合先前树的残差，并组合所有树的预测，以获得最终近似值。

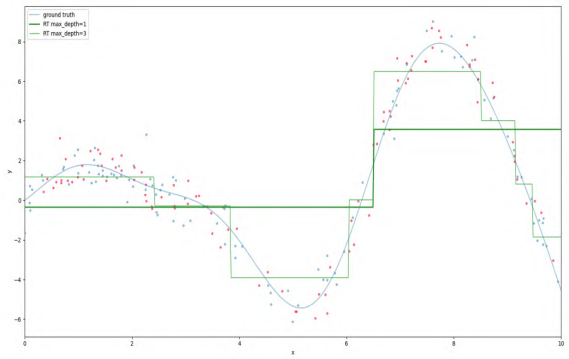

图1　RT max depth=1和RT max depth=2训练结果与ground truth对比

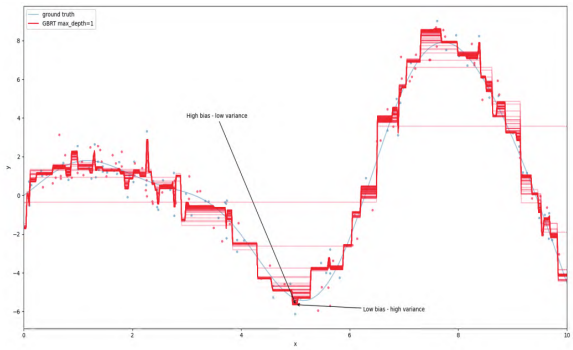

图2　RT depth=1时与ground truth的误差对比

最后，将10棵树添加到具有6个特征的GBRT模型之后，目标变量的近似值将变得更加精确。集合中的每棵树都将在目标变量的分段常数近似中拟合先前树的残差。最终的近似值将是所有树所做预测的组合，这将产生更强大、更准确的模型。

可以看到，图1显示可以防止使用树形结构的过拟合方法来正则化结果；图2显示当RT为1的时存在高误差的情况；图3显示通过正则化，交叉验证等技术可以减少误差，找到最佳数量的树并防止过拟合，确保实证分析的有效性和准确性。

综上所述，可以发现盈利能力和营运效率对中小企业评价模型产生较大的影响，这说明二者在评价企业成长性时发挥了至关重要的作用。同时，信用风险、管理能力和技术创新能力也在一定程度上影响评价结果；而发展潜力对于模型的影响较小，这说明在中小企业成长过程中可以优先解决盈利、营运和信用等对成长影响较为明显的方面，再着重提升发展潜力有助于企业更好地成长。

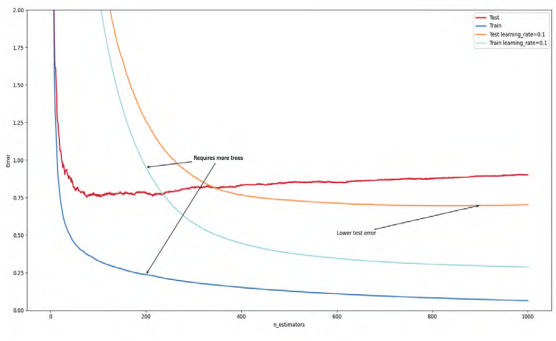

图3　使用stacking算法和Co-training模块后GBRT模型的误差对比

三、结论

中小企业成长过程中的影响因素众多，对其成长性的评价实际上是一个不够准确的考量方式。本文借助已有的评价经验，研究基于GBRT算法的评价模型，该模型的优点在于，可以在数据不足、准确度不够的条件下，扩大信息来源，提高评价分析的可信度。因此，本文所研究的中小企业成长性评价模型具有一定的实用价值，旨在为相关人员提供有益的参考和借鉴。

参考文献

[1] 邬爱其,贾生华.国外企业成长理论研究框架探析[J].外国经济与管理，2002（12）:2-5.

[2] 成璐璐,谢恩,李瑜.关系嵌入视角下政治联系与中小企业研发强度——制度环境与市场竞争的调节作用[J].华东经济管理，2020，34（4）:46-53.

[3] 张倩.基于突变级数法的中小企业成长性评价模型及应用[J].财会通讯:中，2011（11）:56-57.