互联网企业海外并购财务风险大数据预警研究
——基于Stacking集成学习

江乾坤，王成哲
（浙江理工大学 经济管理学院，杭州 310018）

摘 要：数智化时代中国互联网企业海外并购强势崛起但风险巨大，机器学习与非财务信息可为此类风险预警提供新思路。在已有研究基础上，选取2013—2020年45家中国互联网上市公司56起海外并购事件的大数据，构建Stacking集成学习模型进行大数据财务风险预警因子挖掘。结果表明，Stacking集成学习模型相比其他机器学习模型的大数据预警效果更好；运营能力等传统型财务指标依然是互联网企业海外并购财务风险大数据预警的首选指标，但股吧评论等创新型非财务指标也具有重要预警价值。研究结论提供了Stacking机器学习与利益相关者大数据信息有助于预警互联网企业海外并购财务风险的经验证据，为互联网企业、投资者、监管者等进行海外并购财务风险管控决策提供了重要参考。

关键词：Stacking集成学习；海外并购；大数据预警；股吧评论；互联网企业
中图分类号：F272.5 **文献标志码：**A **文章编号：**1002—980X(2023)9—0147—14

一、引言

随着全球数字经济的强劲发展和“数字丝绸之路”倡议的深入推进，以字节跳动、腾讯、阿里巴巴为首的我国互联网企业相继扛起新兴技术产业“出海”的大旗，迅速崛起并在海外并购领域崭露头角。然而，互联网行业作为新兴行业，爆发式成长的背后伴随着高风险。由于互联网企业特有的轻资产结构，资金链紧张成为常态，且其融资方式主要倾向于风险投资和私募，这显著地增加了运营成本和流动性方面的风险，从而引发财务危机。另外，互联网行业的竞争激烈，在“赢者通吃”的市场上只要技术略微突破便可能吸引大批客户，相反技术落后企业便很快会被市场所遗弃（蒋殿春和唐浩丹，2021）。在这种激烈的市场竞争下，互联网企业需要不断创新和提供差异化服务才能生存下来。此外，随着国内人口红利见顶、内需供给增长变缓、智能手机销量下滑，国内互联网各领域增速在逐渐回落，为了追求业务的增长，互联网公司必须要拓展新的市场，“走出去”成为互联网企业的必然选择（郭全中和李祖岳，2023）。然而，在追随互联网巨头“走出去”的过程中，许多新兴互联网企业盲目扩张、过度投资而忽视风险管理，导致内部控制和抗风险能力滞后于扩张速度，造成运营混乱，从而引发财务危机。创造奇迹的同时也暗藏阻碍与风险，例如，暴风影音因为盲目并购英国体育媒体服务公司MPS（MP&Silva）而破产退市，联络互动因为收购美国电商公司Newegg而一度巨亏被特别处理（ST）等。错综复杂的风险因素交织作用于互联网企业海外并购的各个流程，最终效果会以财务指标予以呈现。互联网企业正掀起新一轮国际化投资浪潮，如何应对错综复杂的全球投资环境以避免财务危机？如何利用大数据、云计算、人工智能等新技术进行国际化投资风险预警？如何提升互联网企业跨国并购风险管控能力？因此，有效识别我国互联网企业海外并购财务风险因子，进而制定相应的财务风险预警策略势在必行。

海外并购风险的传统预警手段主要是企业或专业机构的尽职调查、各类机构发布国家投资风险评估报告等单指标、定性、静态模式，在“世界是平的”互联互通时代，这已不能满足风险管控实时决策的需要。本文借助大数据技术开发多指标、定量、动态模型，克服模型设定的片面性和简单性，突破自选择问题，克服了数据不完全性、主观性和时滞性缺陷。

目前财务风险预警模型研究轨迹可分为三个代际：第一代为单一变量分析法；第二代为多元变量和条件

收稿日期：2023-05-09
基金项目：国家社会科学规划课题“‘数字丝绸之路’下互联网企业国际化投资风险的大数据预警与管控研究”（18BGJ013）；浙江理工大学科研启动基金“互联网企业国际化投资风险管控研究”（21092263-Y）
作者简介：江乾坤，博士，浙江理工大学经管学院教授，研究方向：资本市场，海外并购，财务数字化等；王成哲，浙江理工大学经济管理学院硕士研究生，研究方向：财务管理。

概率分析法，如Z分值、逻辑回归模型等；第三代为人工智能分析法，如聚类、随机森林、BP神经网络、支持向量机等（肖毅等，2020）。随着大数据技术的日渐成熟，如何构建机器学习等智能财务风险预警模型正成为新的研究方向。对于财务风险预警因子，现有研究大多局限于战略选择风险、政治风险、融资风险等单一风险或几种风险对互联网企业海外并购的影响，如何引入股吧评论等非财务信息值得期待。

虽然机器学习已广泛运用于风险预警模型构建，但多基于基学习器的单一分类算法和预测，且在实际中仍会遇到诸多难题（杨剑锋等，2019）。本文的贡献在于：首先，已有研究大多集中于“重资产”类的制造型企业海外并购，本文研究对象是聚焦“轻资产”类的互联网企业海外并购，拓展了海外并购风险预警研究；其次，通过大数据证实Stacking集成学习模型相比随机森林（RF）等其他机器学习模型的财务风险预警效果更好；第三，通过Stacking集成学习模型发现，运营能力等传统型财务指标依然是互联网企业海外并购风险预警因子的首选指标，但股吧评论等创新型非财务指标也具有重要的预警价值。

二、文献回顾

现有财务风险预警模型的构建可概括为两个维度（肖毅等，2020）：一是预测方法经历了从单一传统的统计学方法到基于人工智能的机器学习方法的演化；二是风险因子从固定财务比率到通过数据挖掘方法进行数据筛选以选择财务比率，再到引入非财务因素。本文将从智能财务预警模型和互联网企业海外并购风险因子两方面进行梳理。

（一）智能财务风险预警模型

现有智能财务风险预警模型可总结为：①单分类器模型。包括Z分值、Logit、Probit及累积求和模型等统计分析类；人工神经网络、遗传算法、粗糙集、决策树、支持向量机等人工智能类。②混合单分类器。将两个模型串联混合或融合两三种单分类器模型来产生一种新的预测模型。③多分类器组合模型，包括单分类器的并联组合和串联组合（滕晓东和宋国荣，2021）。不过，上述智能财务风险预警研究存在诸多改进之处：一是单分类器模型研究尚不深入；二是多分类器组合模型研究较少；三是忽视专家经验知识和非财务信息对财务风险预警的重要作用；四是针对中国市场开展实证研究的经验证据还不够充分。

当前机器学习算法主要分三类：一是基本分类算法，典型代表是支持向量机（SVM）、逻辑回归（LR）、朴素贝叶斯（Bayes）、邻近算法（KNN）和决策树（DT）；二是神经网络算法（ANN），典型代表是BP神经网络模型和多层感知机（MLP）；三是集成分类算法，典型代表是随机森林（RF）和极端梯度提升（XGBoost）。其中，Gini系数（CART）等决策树算法模型往往会出现过拟合，ANN模型只能高度匹配局部经济状况，模型的大局匹配能力不高，而集成分类算法最为常用。集成学习通过构建并组合优化多个模型来完成学习任务，虽然其得到的也是“弱学习器”，但优点在于可以产生多种“弱学习器”并将它们集成为一个“强学习器”，该新学习器在泛化性能和预测精度方面具有明显的优势。从以往实证结果来看，相比其他机器学习算法，采用集成学习算法对于财务困境企业的预测更为准确（任婷婷等，2021）。

目前最为成熟和发展最壮大的三种集成学习算法（Chowdhury et al，2015）：一是Bagging，包括RF、极端随机树（ET）等，可减少方差；二是Boosting，包括梯度提升算法（Adaboost）、梯度提升决策树（GBDT）和轻量级GBM梯度提升机（LGBM）等，可减少偏差；三是Stacking。三种算法在样本选择、样例权重、预测函数、并行计算、目标侧重上各有千秋。但一般的集成算法是通过某种方式融合多个相同的学习器，而Stacking集成学习策略则更为强大，其通过将多个不同的基本学习器的预测结果作为新的特征输入一个元学习器中，从而获得更准确和泛化能力更强的预测结果（林萍和吕健超，2023）。在Stacking算法中，需要进行两个阶段的学习：第一阶段是使用多个基本学习器对原始数据进行训练和拟合，得到多个基本模型；第二阶段是使用一个元学习器将多个基本模型的预测结果组合起来，生成最终的预测结果。Stacking集成学习方法能够兼顾多个基模型和元模型的学习能力，发挥各模型优势，进一步提高预测精度。此外，由于Stacking集成学习方法及选取模型的自身优势，该模型具有可移植性（李美玉等，2023），在其他应用情境下实现风险预警，例如，信用债违约风险预警（刘晓等，2023）、P2P网贷违约风险预警（丁岚和骆品亮，2017）等。

从智能财务危机预警模型实践来看，通常是先选取财务类指标，包括企业偿债能力、企业盈利能力、企业营运能力、企业现金流量水平、企业发展能力、资本结构（吴春雷和马林梅，2007）。由于财务信息存在滞后

性，应引入多角度的非财务信息，从不同侧面预测企业财务危机的风险源，进而提升预警模型的预测价值（肖毅等，2020），例如，监事总规模、审计意见和创新成长能力、大股东持股比例和独立董事比例（吕峻，2014）、网络舆情（宋彪等，2015）、系统性风险（杨子晖等，2022）、线上运营能力、投诉途径、登陆方式与合作第三方网络平台数量。此外，通过引入新闻媒体和股吧评论等运用大数据分析的指标，财务危机预警模型可以得到有效的改进，从而提高其预警效果，同时减轻传统财务指标的滞后性（宋彪等，2015）。可见，融合大数据与机器学习算法的智能财务风险预警模型不仅可行，而且往往会挖掘很多新型的预警因子。

（二）互联网企业海外并购及其风险因子研究

相比于国内并购而言，跨国并购所涉及的政治、经济、文化等风险问题更为错综复杂（王静，2020），例如东道国媒体负面情绪强烈（晏艳阳和汤会登，2023）；数据风险日益突出（马述忠等，2023）；“来源国劣势”引发东道国政府的监管阻挠（杨勃等，2020）；贸易堡垒带来的跨国并购障碍与风险（杨连星，2021）；文化差异导致并购整合失败（Ahern et al，2015）；制度环境差异大导致并购双方信息不对称（Ahmad et al，2019）、法律风险（俞锋和池仁勇，2015）等宏观因素。但是，这些研究大多限于单一风险或几种风险因子，且大多采用传统实证方法，如Logistic回归等方法，鲜有运用大数据的机器学习方法。而随着大数据技术日益兴盛，通过机器学习模型挖掘更丰富的互联网企业海外并购风险因子已成为可能。

随着国内互联网市场进入存量市场竞争时代，互联网行业“出海”已成趋势，这对以往大多针对于传统制造业的跨国并购研究提出了新的挑战。近年来，部分文献开始对互联网企业国际化展开探索式研究（Vecchi and Brennan，2022；冯乾彬等，2023）。Luo（2021）提出主流的国际化投资理论难以适用于中国互联网行业等新兴行业的投资行为，传统的所有权优势、区位优势和内部化优势在数字经济时代有所削弱。在互联网企业进行跨国并购时，东道国的市场规模、地理距离不再是企业着重考虑的因素，而是更倾向于获取东道国丰富的数字技术和研发资源（蒋殿春和唐浩丹，2021）。相比于传统制造业，互联网行业的敏感性会导致企业在并购时会遭受着更为严厉的东道国政府监管（郭全中和李祖岳，2023），例如，近年来美国对我国的中兴、华为和字节跳动等互联网企业的长臂管辖与定点打击及美国外资投资委员会（CFIUS）以国安理由介入调查并取消的并购案例越来越多。欧盟出台的《通用数据保护条例》（GDPR）等数据隐私法规的出台也对我国互联网企业出海提出了更高的要求（马述忠等，2023）。区别于传统制造业跨国公司，互联网企业独特的成长路径蕴涵着特有的海外并购风险（楼润平等，2019），因此有必要对互联网企业国际化作更深入的探讨。

综合来看，现有文献机器学习研究主体多是上市公司国内并购（王言等，2021），较少专注海外并购事件；研究对象以传统制造业为主，较少专注互联网企业；预警指标体系以微观（企业）财务指标为主（Jia et al，2020），较少涉及宏观（国家）和中观（行业），且对跨层面多角度的影响因素的综合分析较少；预警风险因子以财务类指标为主，非财务类指标已经逐渐增多（陈艺云，2022）；研究方法已经大量探索机器学习模型，但集成学习及Stacking算法模型少见。为此，基于互联网企业海外并购事件及其文献，本文从国家宏观、行业中观、企业微观和大数据4个维度构建互联网企业海外并购财务风险大数据预警指标，通过算法优化构建集成预测模型，并对比不同学习算法在跨国并购风险预警的预测效果，以期为海外并购风险管控提供新思路。

三、Stacking集成学习算法建模与风险因子指标体系构建

（一）研究思路

本文设计的基于Stacking模型的互联网企业海外并购财务风险大数据预警模型实施路线如图1所示。国内和国外并购交易分析平台中记录了大量互联网企业海外并购记录数据，本文首先通过网络爬虫、手工等方法收集我国互联网企业海外并购的样本。除基础数据预处理工作外，本文就可能出现的样本过拟合和特征维度过多的问题提出了解决方案。在模型设计和实施阶段，依据“好而不同”的原则在模型候选列表（包括集成学习模型和非集成学习模型）中进行随机选择并针对海外并购数据集完成训练，并采用机器学习任务中常用的准确率和area under curve（AUC）值等指标进行模型评估，选取预测精度最高的组合模型作为本文的基模型组合。接下来，基于Stacking集成学习的思路，本文对单分类器的输出结果进行特征融合优化，并将其作为输入进行元模型的训练，以输出最终的预测结果。最后，通过输出特征重要性图来分析模型中各个特征对预测结果的影响程度。这有助于理解模型对于不同特征的关注程度，并有助于特征选择和模型调整的优化工作。

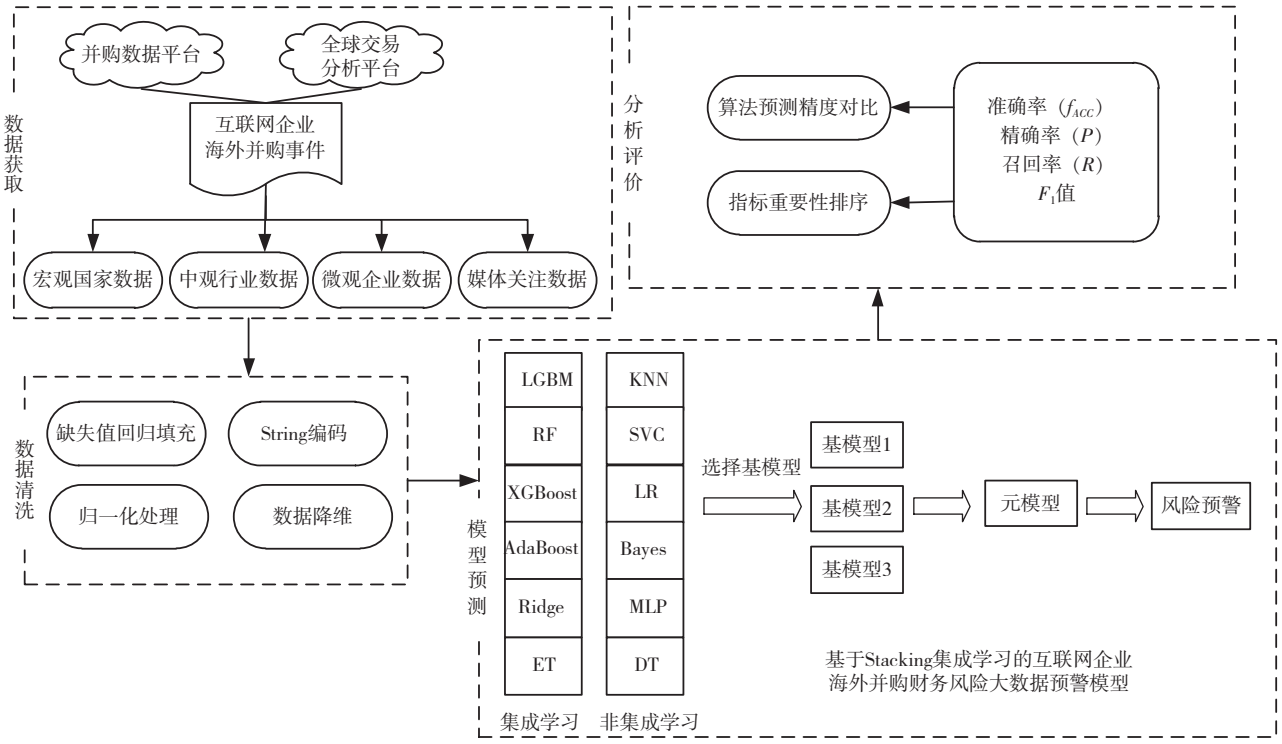


图 1　模型实施路线图

（二）Stacking集成学习算法建模

集成学习技术是将一系列基学习器通过迭代、组合等方式组成新的机器学习模型来降低方差及提高模型的泛化性能（Dasarathy and Sheela，1979），首先，依据预先设定的规则生成多个分类器；其次，利用预设定的组合规则将这些分类器合理地组合起来，形成一个元分类器，其泛化能力更优于单一分类器；最后，综合分析多个分类器的预测结果，得出最终的输出结果。基于“堆叠泛化”（stacked generalization）概念，Wolpert（1992）认为集成学习是一种将多重机器学习模型分类、分层，最后通过一类投票（vote）方法输出模型最终分类结果的算法模型。对比传统的基于投票法的集成学习模型与Stacking模型，后者的分类准确性均优于前者（Georgios et al，2005）。Stacking算法使用特殊的结合方法，可以将不同类型的机器学习算法汇集并堆叠成为一个新的学习器（徐继和杨云，2018）。

Stacking算法建模过程如图2：首先，对数据集进行重采样，获取多个子集，一般分为与基学习器个数相同的份数。第一层学习模型通常是指对原始数据即没有标签的数据进行预测并进行有监督的学习。本文所用的数据均是在已有事实结果的情况下获取，数据已经有了明确结果，故第一层学习模型不再考虑。基学习器是指在构建Stacking算法中用于构建第二层预测模型的机器学习算法。每个基学习器仅使用一个其他基学习器未预测过的子集来作为预测集，以保证这个子集未参与到训练过程之中，且可以减少过拟合程度。通常在选择基学习器时，选择计算方法有偏差的弱学习器来产生分类结果，以免导致后续的训练受第二层结果影响过大，造成结果方差偏离较大。在分配训练子集过程中，应当避免每一块数据索引互相重叠（史佳琪和张建华，2019），以防最终输出结果出现严重的过拟合。

其次，得到所有基学习器的输出结果后，对相互之间的结果进行相关性分析。筛选出相关性较差的输出结果，保留其算法模型，而对于相关度较高的模型则保留预测结果最好的一组模型。其原因在于，不同的算法本质上是不同维度及不同的数据结构角度拟合数据，然后根据不同的原理来建立模型，而最终的叠加是一个纠错过程（徐继和杨云，2018）。这就使得对于整体Stacking集成学习模型而言，基学习器的召回率比准确率更重要。本文选择Pearson法来衡量各个模型的差异程度，其计算方法如式（1）所示。

$$r_{xy}=\frac{\sum_{i=1}^{m}(x-\bar{x})(y-\bar{y})}{\sqrt{\sum_{i=1}^{m}(x-\bar{x})^2}\sqrt{\sum_{i=1}^{m}(y-\bar{y})^2}} \tag{1}$$

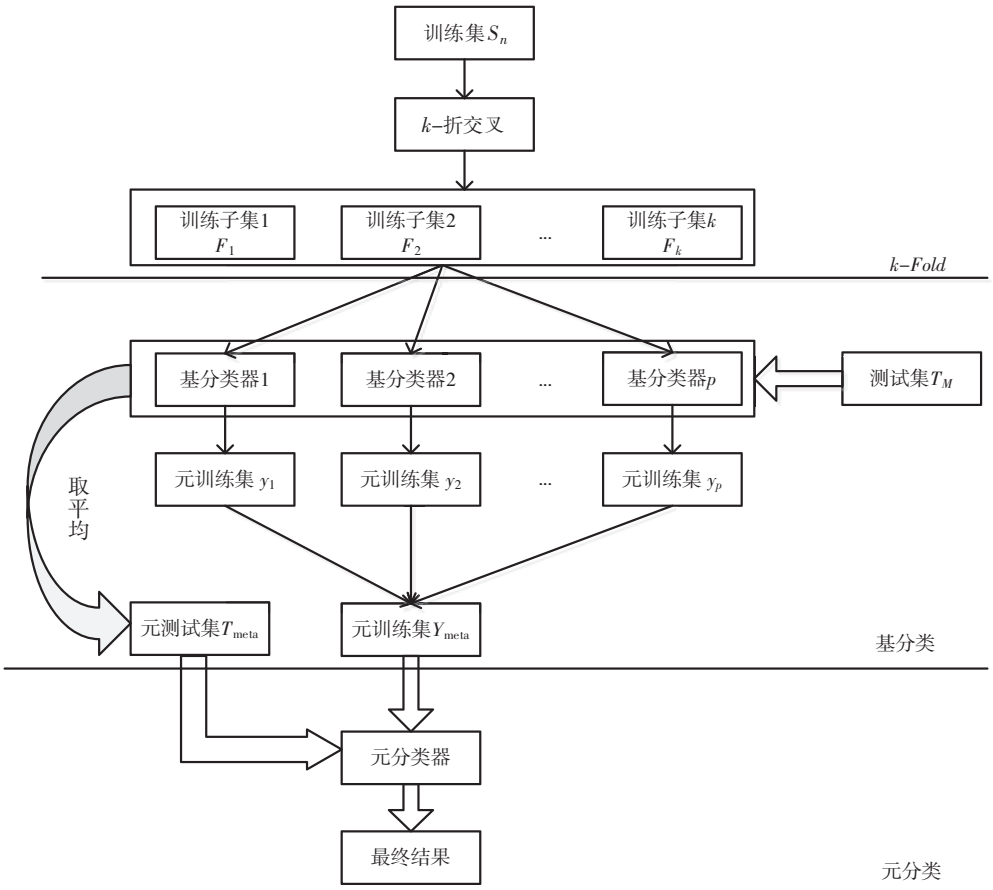


图 2　Stacking算法建模过程

其中：x 和 y 分别为两不同模型输出的预测值；i 为观测值，共 m 组预测值。r_{xy} 越小，模型匹配度越高。再者，选择所有相关性较差的结果组合记录其所对应的基学习器，得到第二层基学习器集合。这些基学习器在最终的集成学习算法之中将会反复训练堆叠，结果也会不断做交叉验证，最终选取得到精确率更高的集成学习模型。

最后，第三层通常选用投票法来产生最终的预测结果。基于陈铁明和马继霞（2012）等已有研究，通常赋予最优模型以更高的权重。根据随机森林或其他树形决策分类器的特征，在已经获得数据分类结果的情况下，可以使预测结果更好的模型得到更高权重，也可以使用加权投票法来简化算法流程（徐继伟和杨云，2018）：$H(x)=\sum_{i=i}^{T} w_i h_i(x)$，其中 w_i 为第 i 个个体学习器的权值；$h_i(x)$ 为第 i 个学习器（共 T 个）的误差，通常 $w_i>0$ 且 $\sum_{i=1}^{T} w_i>0$。或可以采用平均法：$H(x)=\frac{1}{T}\sum_{i=1}^{T} h_i(x)$，其中 i 为第 i 个学习器。Stacking算法具体表示如下：对于一个样本集合 $D=\left\{(x_p, y_p), p=1,2,3,\cdots,N\right\}$，$y_p$ 是第 p 个样本的结果，x_p 为第 p 个样本所对应的特征集。

（三）互联网企业海外并购财务风险预警因子体系

1. 互联网企业海外并购风险预警因子体系

综合现有研究，本文构建的中国互联网企业海外并购风险预警因子包括4个维度（表1），共计86个指标。其中，股吧评论属于大数据非财务指标，下文将详细解析，其他类指标限于篇幅不再详析。这些风险预警因子相对独立又相互关联，从风险演化链角度来看，东道国宏观风险因子、市场中观风险因子、企业微观风险因子往往会依次显现，媒体关注等大数据预警因子则凭借独特的实时动态优势贯穿其中，它们的综合预警效果最终会通过主并企业财务危机形式呈现，而上述纷繁复杂的风险预警过程无法采用传统的财务风险预警模型，需要引入以集成学习为代表的的智能财务危机预警模型。

表 1　互联网企业海外并购财务风险预警因子体系

宏观：国家风险	政治风险		X1 执政时间；X2 政府稳定性；X3 军事干预政治；X4 腐败；X5 民主问责；X6 政府有效性；X7 法制；X8 外部冲突
	经济基础		X9 市场规模；X10 发展水平；X11 经济增速；X12 经济波动性；X13 通货膨胀率；X14 贸易开放度；X15 失业率；X16 投资开放度；X17 资本账户开放度；X18 收入分配
	偿债能力		X19 外债占 GDP 比重；X20 财政余额占 GDP 比重；X21 公共债务占 GDP 比重；X22 经常账户余额占 GDP 比重；X23 短期外债占总外债比重；X24 外债占外汇储备比重；X25 银行业不良资产比重；X26 贸易条件；X27 是否为储备货币发行国
	社会弹性		X28 内部冲突；X29 环境政策；X30 资本和人员流动的限制；X31 劳动力市场管制；X32 商业管制；X33 社会安全；X34 教育水平；X35 其他投资风险
	对华关系		X36 是否签订双边投资保护协定（BIT）；X37 投资受阻程度；X38 双边政治关系；X39 贸易依存度；X40 投资依存度；X41 免签情况
中观：市场	市场环境		X42 营商环境（全球营商环境排名）；X43 赫芬达尔指数（市场集中度）
	市场反应		X44 股价波动率（投资者认可度）；X45 评级结论（证券分析师预测）
微观：企业	并购流程	并购战略	X46 行业选择；X47 区位选择
		并购交易	X48 并购溢价率（并购定价风险）；X49 支付方式
		并购整合	X50 整合方式；X51 商誉减值（并购整合效果判断）
	财务能力	偿债能力	X52 资产负债率；X53 带息负债比率；X54 速动比率；X55 现金流动负债比率；X56 经营活动产生的现金流量净额/带息债务
		盈利能力	X57 净资产收益率；X58 总资产报酬率；X59 主营业务利润率；X60 成本费用利润率
		营运能力	X61 总资产周转率；X62 应收帐款周转率；X63 流动资产周转率；X64 存货周转率
		发展能力	X65 总资产增长率；X66 销售利润增长率；X67 资本保值增值率
		现金流量	X68 净利润现金净含量；X69 营业利润现金净含量；X70 营业收入现金净含量；X71 全部现金回收率；X72 营运指数；X73 每股经营活动现金流量净额；X74 每股企业自由现金流量
	技术创新投入		X75 技术投入比率（企业本年科技支出/本年营业收入）
	内部控制质量		X76 内部控制衡量指标（DIB 迪博数据库）
	审计意见		X77 审计意见类型
媒体关注（大数据）	网络搜索		A1 百度指数（百度搜索词条）
	股吧评论		B1：T-2 年贴子数；B2：T-2 年评论数；B3：T-2 年舆论热度；B4：T-2 年积极情绪指数；B5：T-3 年贴子数；B6：T-3 年评论数；B7：T-3 年舆论热度；B8：T-3 年积极情绪指数；第 T 年为主并互联网企业实施海外并购后被标为 ST 的年份，T-2 年、T-3 年表示互联网企业海外并购后被 ST 的前 2 年、前 3 年

注：宏观国家层面的指标的选取参考由中国社会科学院发布的《中国海外投资国家风险评级报告》，指标详情也可参考此报告。

2. 股吧评论指标

从在线信息获取的企业相关大数据，其内容可包含导致企业财务危机方方面面的因素，甚至包含人们尚未认识到的危机根源。在众多网络平台中，股吧平台最为活跃，也是最具有研究价值的平台，股吧平台是媒体、机构投资者、小众投资者、供应商及基金经理之间信息传递的重要媒介，其产生的大数据对于研究公司股票价格和财务状况的变化极具价值（Lai，2022）。股吧评论中不乏资深网民与相关专家对海外并购事件的真知灼见，其言论具有一定的专业性和科学性。它们所传递的信息及情感交流的互动和波动在一定程度上能够反映企业在实施海外并购后的经营及财务状况，因此对企业的财务危机具有一定的预警价值。此外，互联网上的网民对企业的相关行为也会产生反应，这涵盖了线下接触企业的人们所产生的各种情绪。所有这些信息通过线下行为映射到互联网，并通过聚集、排斥和融合的作用在互联网中形成股民情绪，进而形成与相关企业相关的网络舆情（宋彪等，2015）。这些客观、科学的数据可以为财务危机预警提供帮助。不仅大数据与企业财务状况密切相关，而且通过计算机自然语言处理技术进行量化处理，结果更加客观，因此通过大数据量化处理形成的指标可以解决以往非财务指标片面、主观、难以量化的问题。通过分析和监测这些数据，可及早发现潜在的财务风险因素和市场反应，帮助企业及时采取措施避免危机的发生或减轻其影响（段珊珊和朱建明，2016）。

关于股吧评论的指标获取，本文采用 Python 作为编程基础，选取中国最大的财经网站东方财富网作为数据来源，从中批量爬取评论的标题、内容文本、时间等。为了对所爬取的内容文本进行情感分析，本文采用了多个情感词典来构建情感词库，其中包括如下词典：第一，基础词典，主要以知网 HowNet 情感词典为主；第二，网络语言词典，以 BosonNLP 和 SnowNLP 情感词典为主；第三，金融专业领域词典，以证券和财经领域词汇为主；第四，新闻词典，主要以新闻、政策中隐性情感倾向的词汇为主。基于以上的情感词典，加入其他手动搜集的情感词和股吧情感词典（表略），得到本文进行集成学习的评论数据情感词典。此外，在日常交流中，除了情感词典中的积极词汇和消极词汇以外，大量的副词和否定词也经常被用来加强或减弱所要表达的内容。为了更准确地评估文本情感，本文参考 HowNet 情感词典、相关研究和人工收集的信息，整理出副词和否定词的词典（表略），并将它们分为 7 个等级，根据现有的文本情感分析文献进行具体赋值。积极词汇赋值为 1，消极词汇赋值为 -1，副词和否定词的值在 -1.0~2.5，绝对值越高表示程度越强。

另外，根据情感词典和机器学习程序分析股吧评论的情感值。使用jieba分词将爬取到的文本内容的句子分割成词汇，将分割后词语中的情感词与情感词典中的词汇自动进行对比，并使用程度副词进行加权计算得到情感值。之后，根据文本中各词汇的情感值，相加汇总后可得到每一个帖子中文本的情感值。若情感值大于0，则当前主题帖为积极评论贴；若情感值小于0，则当前主题帖为消极评论贴；若情感值为0，则将其定义为中立评论贴。

本文将主并互联网企业实施海外并购后被ST的年份定义为 T 年，由于财务报告发布具有滞后性，T-2年财务数据已包含企业发生财务危机的主要特征——亏损，但这些评论属于在 T-2年财务数据发布之前的评价，并不会夸大财务危机预警效果。因此，本文在考虑大数据指标时，选取企业 T-2 和 T-3 年的股吧平台数据进行观察和分析。

由此，本文给出股吧评论大数据指标的定义见表2。

表2　互联网企业海外并购事件的股吧评论大数据指标

符号	定义	说明
*B*1	*T*-2 年帖子数	ln(*T*-2 年帖子汇总数)
*B*2	*T*-2 年评论数	ln(*T*-2 年帖子下的评论汇总数)
*B*3	*T*-2 年舆论热度	ln(*T*-2 年帖子汇总数+帖子下的评论汇总数)
*B*4	*T*-2 年积极情绪指数	(*T*-2 年积极评论帖子数-*T*-2 年消极评论帖子数)/(*T*-2 年积极评论帖子数+*T*-2 年消极评论帖子数)
*B*5	*T*-3 年帖子数	ln(*T*-3 年帖子汇总数)
*B*6	*T*-3 年评论数	ln(*T*-3 年帖子下的评论汇总数)
*B*7	*T*-3 年舆论热度	ln(*T*-3 年帖子汇总数+帖子下的评论汇总数)
*B*8	*T*-3 年积极情绪指数	(*T*-3 年积极评论帖子数-*T*-3 年消极评论帖子数)/(*T*-3 年积极评论帖子数+ *T*-3 年消极评论帖子数)

资料来源：东方财富网。

综上，本文选取 T-2年帖子数、T-2年评论数、T-2年舆论热度、T-2年积极情绪指数、T-3年帖子数、T-3年评论数、T-3年舆论热度、T-3年积极情绪指数8个大数据指标作为待选项，用于后续的海外并购财务危机预警模型，并借助Python收集到约60万条数据，整理汇总成45家互联网上市公司的56起海外并购事件的8个大数据指标。本文的互联网企业海外并购财务风险大数据预警模型如图3所示。

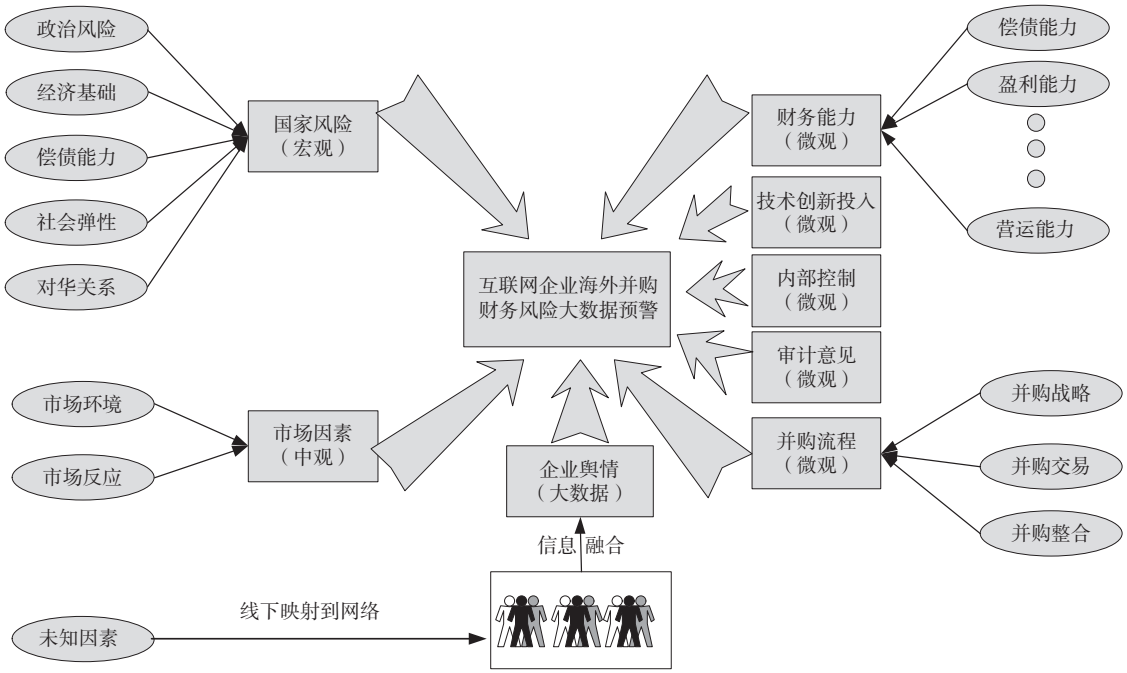


图3　互联网企业海外并购财务风险大数据预警模型

四、互联网企业海外并购财务风险预警的Stacking模型及其数据分析

（一）数据获取

完整且质量高的数据是机器学习的重要基石。在数据获取的过程中，应以数据质量评估标准为导向，以确保数据的完整性、一致性和准确性，并在最终结果的形成中予以体现。同时，需要摒弃传统的逻辑思维方式，不再仅从因果逻辑的角度出发，寻找与实验目标有可能相关联的因素，而是应该尽可能从多个维度收集企业的所有相关信息。这些特征值有可能是以非线性的形式呈现在最终的分类结果之中，传统的线性回归

方式对非线性叠加的特征利用率较差，而借助集成学习算法，则有可能从海量的、杂乱无章且不清晰的数据中找寻到蕴含有规律、有价值和能够理解应用的特征。

1. 数据来源

鉴于许多中国互联网企业注册于开曼群岛等避税天堂，本文以实际营业地或办事机构所在地处于中国大陆并在沪深两市、香港联交所、美国纳斯达克交易所、美国纽约交易所等上市的中国互联网企业为主并企业，以实际营业地或办事机构所在地处于中国大陆以外（不含港澳台地区）的企业为目标企业，以2013年1月1日—2020年12月31日发生的45家中国互联网企业56起海外并购事件为研究样本，具体信息主要源于清科研究中心、Zephyr全球并购交易分析库、中国全球投资跟踪报告（美国企业研究所和传统基金会）、国泰安“海外直接投资”数据库，同花顺iFind等数据平台，通过网络爬虫、并购数据库、手工等方法收集，结合新浪财经、巨潮资讯、东方财富网等多方平台加以验证和筛选，并对如下样本进行剔除：①未对外公告的并购事件；②并购前为ST类公司；③目标公司所在地为港澳台地区、开曼群岛、英属维尔京群岛等避税区；④数据缺失的样本。最终得到56起中国互联网企业跨国并购事件样本。宏观层面的东道国国家风险指标和数据主要来自中国社会科学院世界经济与政治研究所（IIS）发布的历年《中国海外投资国家风险评级报告（2013—2021）》（CROIC-IWEP）；中观层面的数据主要来自于世界银行、百度搜索和同花顺；微观层面的企业数据主要来自国泰安、新浪财经、巨潮资讯等。

2. 数据爬取与存储

为了获取建模所需要的互联网企业海外并购风险因子数据，研究团队编写了爬虫程序，在公开的海外并购相关数据平台上爬取互联网企业海外并购事件的各个维度信息。这些数据平台覆盖了清科研究中心（数据库-并购事件）、新浪财经、巨潮资讯、东方财富网等网站。具体流程如下：①获取链接。根据网站自身统一资源定位符（URL）规则获取各个数据的链接，设置baseURL变量遍历所有数据。②获取信息。利用BeautifulSoup库对html重构成文档树，并加入异常捕获、日志记录增强爬取过程程序的健壮性。随机挂起程序，以减轻网站访问压力。③数据存储。利用轻量级的sqlite3数据库实时存储爬取到的数据。

3. 数据预处理

（1）缺失值回归填充。在原始测试集中，除去对年份进行检索补全之外，发现缺失值分布较为均匀。考虑到数据有部分分布不均衡，本文将测试集中约15%的空缺数据删除，以减少对最终结果的影响。在增添的特征方面，由于对数据的除法运算会出现除无意义（0ERROR），将这一部分跳过之后会出现空值，所以选择回归填充缺失值的方法，分*flag*=0和*flag*=1的情况执行随机森林决策树回归填充缺失。上述缺失值填充原理是：在填补每个特征时，将其他特征的缺失值用0代替，每完成一次回归预测，就将预测值放到原特征矩阵中，再继续填补下一个特征。随着每个特征的填补，有缺失值的特征数量会逐渐减少，每次循环后需要用0填补的特征也会越来越少。当遍历到最后一个特征时，所有其他特征都已经用回归填补了大量有效信息，可以用这些信息来填补缺失最多的特征。最终，遍历所有特征后，数据将不再存在缺失值。

（2）String编码。由于区域特征比较少，对此部分的特征考虑选用独热编码或直接编码。在初步选用的模型尝试后发现直接编码效果比较好，最终采取了直接编码的形式。

（3）归一化处理。由于参数变化范围较大，最终可能会对模型产生影响，需要移除掉名称、区域、行业等不需要标准化的数据后再对其他数据进行归一化处理，将该类数据原始值x使用z-score标准化到x'。数据标准化过程中对序列$x_1, x_2, \cdots, x_n$进行如下变换：$y_i = \dfrac{x_i - \bar{x}}{s}$，其中，$\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$，$s = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$则新序列$y_1, y_2, \cdots, y_n$的均值为0，方差为1，且无量纲。

（4）数据降维。数据降维就是通过特征选择或特征变换操作将数据从原始的D维空间投影到新的K维空间。数据降维方法主要分为两类：一是特征选择，它是在所有的特征中通过子集搜索算法寻找和模型最相关的特征子集的过程，即在所有特征中选择和目标最相关的一些特征，丢弃掉一些不太重要的特征。特征选择可细分为三个类型：①过滤式，即根据特征的统计学特性选择特征，例如Relieff算法等；②包裹式，即通过训练机器学习模型来选择特征，例如支持向量机递归特征消除（SVM-RFE）方法等；③嵌入式，即在训练机器学习模型的同时选择了特征，例如逻辑回归、LASSO回归（最小绝对值收敛和选择算子算法）。二是特征抽

取，亦称特征降维，它是指通过某种线性变换或非线性变换，将数据从高维空间映射到低维空间，例如主成分分析法（PCA）（Tharwat，2016）。特征选择的数据降维方法符合本文研究目的，同时 Least absolute shrinkage and selection operator（LASSO）回归对于数据的要求极低，能够进行变量筛选和降低模型复杂度。变量筛选是为了在模型拟合过程中选取最重要的变量，从而提高模型的性能和泛化能力。而复杂度调整则是为了避免过拟合现象，即过度拟合训练数据集，而导致在新的数据集上表现不佳的情况。因此，本文选择 LASSO 回归方法进行数据降维。LASSO 回归通过 $L1$ 正则化对回归系数进行惩罚，可以将不重要的变量系数缩小甚至置为 0，从而实现变量筛选和模型复杂度调整。因此，LASSO 回归是一种非常有效的数据降维方法，适用于高维数据的建模和特征选择（Tibshirani，1996）。变量系数 $\hat{\beta}$ 的计算公式为：$\hat{\beta}_{\text{lasso}} = \text{argmin}\left[\sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right]$。其中，$y_i$ 为第 i 个样本（共 n 个），β_j 为第 j 个参数（共 p 个）。λ（大于 0 的正数）作为调和参数，调节惩罚项（公式后半部分）权重。当 λ 越来越大时，惩罚项的作用将越来越强，模型的大部分回归系数会被约束为 0，因此可以通过控制 λ 来控制所选变量个数。

（二）过拟合问题

Stacking 算法是一类多重算法堆叠而成的强学习器，如同大多数强学习器一样，它容易产生过拟合问题。不过，在构建模型并检验的过程中，模型最终的结果并不是适配训练数据，而是要适配验证数据。本文对 Stacking 建模过程中可能出现过拟合的情况作出如下说明：①如果所获取数据是原始数据，即没有分类完成，需要有监督的学习并完成分类结果，则第一层训练模型的选择中应当避免选择可能出现低方差、高偏差的模型，通常是指强学习器。第一层训练结果通常会作为初始训练集和测试集，使用低方差的模型有可能使最终模型输出一个偏差极大的结果。如果多次对模型进行调整后仍得不到理想的提升，则有可能是第一层训练模型过拟合。②在第二层训练模型的选择之中，除应当选择输出结果相关度较低的基学习器组合之外，还应当注意这部分的训练集拆分不能使得不同的基学习器使用相同的训练集，这会导致训练集和测试集有交叉，影响真实的模型精确率，导致最终输出模型拟合度过高。③在进行特征工程时，如果使用多个特征进行运算得到一个新特征，新特征的使用会显著增强参与运算的特征在模型之中的权重。即使得到的实验数据精确度更高，也要防范过拟合的风险。

（三）实验过程及结果分析

1. 数据处理

本文通过并购数据库、网络爬虫等多种数据渠道共获取了 2013—2020 年 45 家中国互联网上市公司 56 起海外并购事件数据，并购标的涉及 16 个国家及地区，包含 86 个数据维度，并购信息、风险因子，对应数据处理方法分别为编码、归一化、One-Hot 编码等。

2. 样本选取

关于研究样本的分类，本文采用上市公司是否被 ST 作为财务困境的判别标准，ST 公司界定为财务困境公司，非 ST 公司界定为财务健康公司。从样本公司实施海外并购后财务状况可知（表 3）：财务健康公司为 32 家，财务困境公司为 13 家，两者比例约为 2.5∶1。现有研究对智能财务危机预测时，大多将测试样本组和训练样本组的比例设为 1∶2（滕晓东和宋国荣，2021）。遵循这一原则，本文从总研究样本中随机抽取 35% 作为测试样本组，剩下 65% 作为训练样本组。因此，最终的训练样本组由 29 家公司组成，其中财务危机公司 8 家，正常公司 21 家；测试样本组由 16 家公司组成，其中财务危机公司 5 家，正常公司 11 家。

表 3　研究样本概况

样本分组	样本类型	样本个数（个）		样本占比（约）	
测试样本	财务危机公司	5	16	31%（测试组内占比）	35%（总体占比）
	正常公司	11		69%（测试组内占比）	
训练样本	财务危机公司	8	29	28%（训练组内占比）	65%（总体占比）
	正常公司	21		72%（训练组内占比）	

3. 数据降维

经初步处理后，本文通过 LASSO 回归筛选指标，在上述 86 个指标中剔除了系数为 0 的指标，从中筛选出与财务危机预警较为相关的 31 个主要指标作为后续变量，具体变量见表 4。

4. Pearson 相关性分析

Stacking 算法集成多种机器学习算法堆叠成为新的学习器，通过投票法或加权投票等方法来修正基学习

器的错误分类。因此,在选择基学习器时要尽可能选择不同种类的学习器,这可以根据预测结果的二维 Pearson 相关系数作为参考依据。本文在计算后选取了逻辑回归(logistic regression)、岭回归(ridge regression)、极端梯级提升树分类器(XGBoost classifier)、LGBM 分类器(LGBM classifier)及随机森林分类器(random forest classifier)作为基学习器,各类算法的误差 Pearson 相关性分析的热力图见图 4。由图 4 可知,除了 XGBoost 与 Ridge 算法所输出的预测结果相关性较强外,其他算法所输出的预测结果相关性并不明显。因此,可以将这些算法作为基学习器组成最终的 Stacking 算法。

5. 模型训练

将训练集根据基学习器数量进行 k 折交叉(本文 k=5)后得到训练子集。分别使用 sklean 库中 5 种基学习器 LR、Ridge、XGBoost、LGBM 和 RF 算法来训练得到训练模型。

6. 模型质量的评价指标

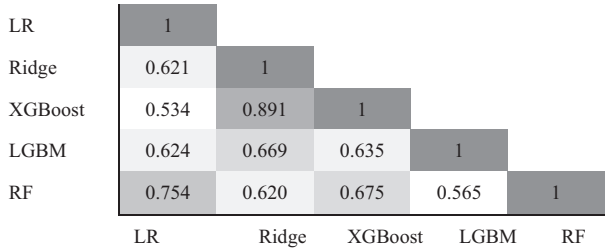
机器学习需要建立模型来解决具体问题,通常需要使用一些指标来评估模型的性能和泛化能力,常用的模型评价指标例如准确率、精确率、召回率、F_1 等,而它们都建立在混淆矩阵(confusion matrix)的基础上。

(1)混淆矩阵。混淆矩阵又被称为错误矩阵,被用来呈现算法性能的可视化效果,通常是监督学习。表 5 中,每一列代表预测值,每一行代表的是实际的类别,其中:*TP* 代表将正例正确识别成正例的数量;*FP* 代表将反例错误识别成正例的数量;*FN* 代表将正例错误识别成反例的数量;*TN* 代表将反例正确识别成反例的数量。

(2)评价指标。通过混淆矩阵,可以得到模型的准确率等指标,具体评价指标解释见表 6。

表 4　互联网企业海外并购财务风险预警指标(含大数据)

宏观:国家风险	政治风险		*X*2 政府稳定性
	经济基础		*X*11 经济增速;*X*14 投资开放度
	偿债能力		*X*20 财政余额/GDP;*X*26 贸易条件
	社会弹性		*X*32 商业管制
	对华关系		*X*38 双边政治关系;*X*40 投资依存度
中观:市场	市场环境		*X*42 营商环境;*X*43 赫芬达尔指数
	市场反应		*X*44 股价波动率
微观:企业	并购战略		*X*47 区位选择
	并购交易		*X*48 并购溢价率
	并购整合		*X*51 商誉减值
	财务能力	负债能力	*X*52 资产负债率;*X*54 速动比率;*X*56 经营活动现金流量净额/带息债务
		盈利能力	*X*58 总资产报酬率
		营运能力	*X*61 总资产周转率;*X*64 存货周转率
		发展能力	*X*66 总资产增长率
		现金流量	*X*69 营业利润现金净含量;*X*70 营业收入现金净含量
	技术创新投入		*X*75 技术投入比率
	内部控制质量		*X*76 内部控制质量
	审计意见		*X*77 审计意见类型
媒体关注(大数据)	新闻媒体		*A*1 百度搜索词条
	股吧评论		*T*-3 年积极情绪指数、*T*-3 年评论数、*T*-2 年积极情绪指数、*T*-2 年帖子数



其中数据值的大小以颜色来进行区分

图 4　各类机器学习算法误差 Pearson 相关性分析的热力图

表 5　混淆矩阵

混淆矩阵		预测结果	
		预测为正例	预测为反例
真实结果	实际正例	*TP*(TruePositive)	*FN*(FalseNegative)
	实际反例	*FP*(FalsePositive)	*TN*(TrueNegative)

表 6　机器学习模型评价指标

指标名称	计算公式	指标解释
准确率(f_{ACC})	$(TP+TN)/(TP+FP+FN+TN)$	预测正确的样本占总观测数比重,用以衡量模型的整体效果
精确率(P)	$TP/(TP+FP)$	在所有被预测为正例的样本中,真正为正例的样本所占的比例,代表着对正样本结果中的预测准确程度,精确率越高,说明模型在预测正例时的准确性越高
召回率(R)	$TP/(TP+FN)$	在所有实际为正例的样本中,被正确预测为正例的样本所占的比例,召回率越高,说明模型在检测正例时的能力越强
F_1	$2PR/(P+R)$	F_1 综合考虑了精确率和召回率,可以用来综合评估模型的性能。值越大,输出结果越好

7. 堆叠次数

基学习器参数设置无需过于苛刻追求精度,这是由 Stacking 堆叠算法的计算原理决定的。对基学习器进行五轮迭代后投票,分别输出每轮堆叠的精确率、准确率和召回率。选择精确率最高的一组参数并得到最终 Stacking 模型的输出结果。图 5 为五轮迭代过程中精确率、召回率和准确率的变化。最终结果使用准确率来进行比对分析,设置不同次数堆叠,会对结果产生细微影响。本文自第零次堆叠开始总计最高堆叠 6 次,得到了七种结果(图 5):其中,精确率和准确率均以检出海外并购风险互联网企业数量为分子。由计算结果

可以看出，整体的准确率随着堆叠次数的增加呈非线性变化。在研究中，需要根据实际情况参考不同的指标。本文希望系统能尽量全面的检出含有海外并购财务风险的互联网企业，因此召回率和准确率是本文的主要参考指标。通过对比，本文选取的堆叠次数为1。

8. 输出结果

各模型的评价指标输出结果见表7。对比传统的机器学习结果，Stacking模型能够获取更高的准确率(93.4%)，召回率(95.5%)也达到最高，说明本模型能够最大限度检出互联网企业海外并购后当前是否有可能处于风险状况；Stacking模型的F_1(86.2)高于其他模型的F_1，说明其稳健性较其他模型更为突出。因此，Stacking集成学习得到的相关指标证明该模型的可靠性，可以用于对互联网企业海外并购财务风险的预警。

从测试样本的Stacking模型预测结果来看(表8)，正常企业与ST企业的预测正确率分别为90.9%和80.0%，预测效果良好。

9. 预警指标

Stacking模型无法得到一个简单的数学公式来表示预测结果，属于“黑盒子”预测，因为它是通过多个基模型和一个次级模型的组合来得到预测结果的。但是，本文可以通过输出特征重要性图来分析模型中各个特征对预测结果的影响程度，以帮助理解模型对于各个特征的关注程度，帮助优化特征选择和模型调整。

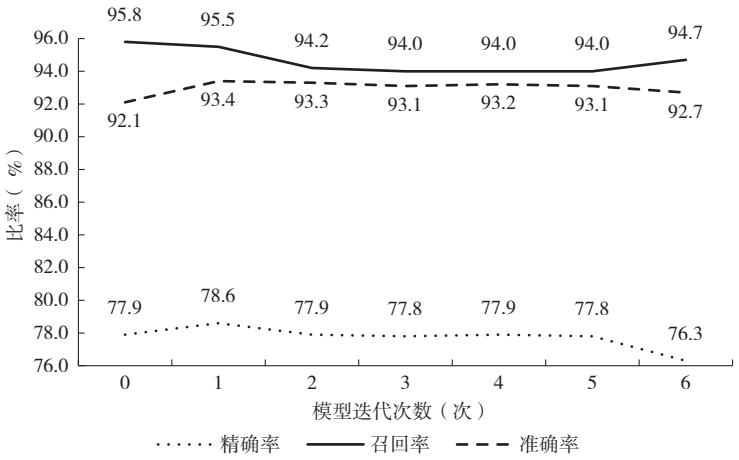


图5 模型迭代次数与精确率、召回率和准确率的关系

表7 各模型性能度量指标值

模型名称	准确率(f_{ACC})(%)	精确率(P)(%)	召回率(R)(%)	F_1
LR	85.1	80.1	77.4	78.7
Ridge	71.7	79.4	75.2	77.2
XGB	82.3	69.8	80.1	74.6
LGBM	80.3	77.3	83.9	80.5
RF	91.4	81.1	90.3	85.5
Stacking	93.4	78.6	95.5	86.2

表8 Stacking模型预测结果(测试样本16个)

企业类型	正常企业(预测)	财务危机企业(预测)	预测正确率(%)
正常企业(实际)	10	1	90.9
财务危机企业(实际)	1	4	80.0

(1)重要性排序。根据Stacking的重要性分析，得到有利于财务危机预警的15个重要指标如图6所示。其中，基于国家风险维度是“投资开放度”指标，基于市场风险维度是“股价波动率”指标，基于财务能力维度是“总资产周转率”“营业收入现金净含量”“总资产报酬率”“经营活动产生的现金流量净额/带息债务”“存货

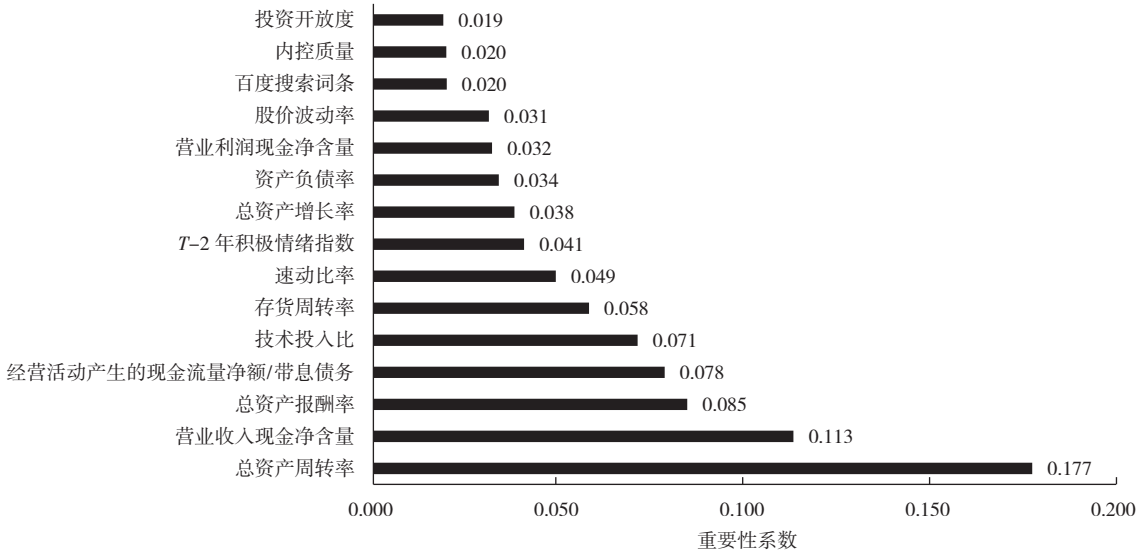


图6 Stacking模型显示的前15个预警指标

周转率”“速动比率”“营业利润现金净含量”“总资产增长率”“资产负债率”这些指标，基于技术创新维度是“技术投入比率”指标，基于内部控制维度的是“内部控制质量”指标，基于大数据维度是“百度搜索词条”指标和“*T*-2积极情绪指数”指标。可见，目前影响互联网企业国际化投资风险的预警指标主要是微观层面，主并企业的财务能力指标有9个，且从重要性排序来看，除了“技术投入比率”指标外，它们占据前10位。其次是技术创新投入（技术投入比率）和股吧评论（*T*-2积极情绪指数），它们分列第5、11位。最后是市场反应（股价波动率）与新闻媒体（百度搜索词条），它们分列第12、13位；最后是内部控制（内部控制质量）与经济基础（投资开放度）。

（2）进一步研究。如果把预警指标重新分类，财务预警指标视为传统类，非财务预警指标视为创新类，对Stacking模型预测结果进一步分析，且将输入模型的特征进行重要性排序，可以得到两类新的互联网企业海外并购风险的预警风险因子（图7、图8）。其中，排名前五的传统型财务预警指标分别是：总资产周转率、营业收入现金净含量、总资产报酬率、经营活动产生的现金流量净额/带息债务、技术投入比。具体来说，以企业营运能力指标（总资产周转率和流动资产周转率）为主，其次是企业盈利能力指标（总资产报酬率），然后是企业负债能力指标（经营活动产生的现金流量净额/带息债务）、企业创新能力（技术投入比）。排名前五的创新型非财务预警指标分别是：*T*-2年积极情绪指数、股价波动率、百度搜索词条、内部控制质量、投资开放度，它们分别反映了投资者关注、股价走势、网络搜索、企业内控质量和东道国经济基础对中国互联网企业海外并购风险具有一定的预警价值。

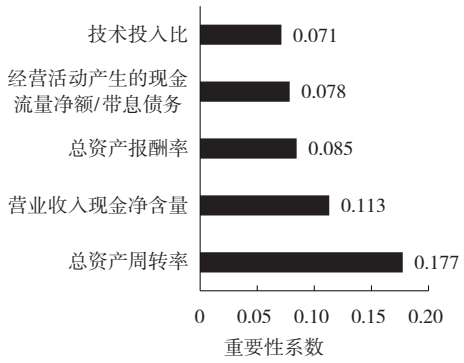


图7　Stacking模型显示的前5个财务预警指标

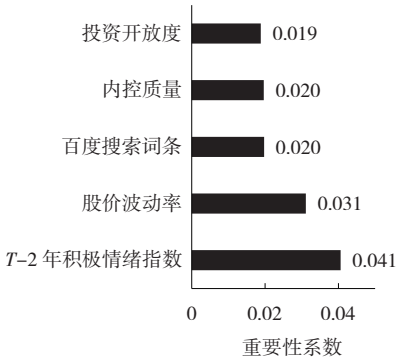


图8　Stacking模型显示的前5个创新型非财务预警指标

综合来看，企业营运能力、现金流量、盈利能力、负债能力和技术创新等传统型财务指标依然是互联网企业海外并购风险预警的首选指标，但是股吧评论、股价波动率、网络搜索、企业内控质量与东道国投资开放度等创新型非财务指标对互联网企业海外并购风险预警也具有重要的参考价值。

五、结论

数智化时代，机器学习方法与股吧评论等大数据信息为互联网企业海外并购风险预警提供了新的思路。本文基于45家中国互联网企业海外并购样本及其86个风险预警指标，通过Stacking集成学习模型进行机器学习，研究发现：相对于LR、Ridge、XGBoost、LGBM、RF等机器学习模型，Stacking集成学习模型的财务风险预警效果更好；关于互联网企业海外并购风险预警因子的选择，企业营运能力、现金流量、盈利能力、负债能力和技术创新等传统型财务指标依然是首选指标，但股吧评论、股价波动率、网络搜索、企业内控质量与东道国投资开放度等创新型非财务指标也具有重要的预警价值。

本文的不足之处在于，一是互联网企业海外并购研究样本只有45家，在划分为训练组与测试组后，测试组样本数量偏少；二是大数据维度的预警因子偏少，只涵盖新闻媒体与股吧评论。下一步研究将加大样本数量，纳入更多的大数据预警因子，例如，东道国媒体舆论，上市企业年报管理者陈述语调等，且深入探讨预警因子与互联网企业海外并购风险之间的因果关系，为互联网企业海外并购风险管控提供更多的决策参考。

参考文献

[1] 陈铁明，马继霞，2012. 一种新的快速特征选择和数据分类方法[J]. 计算机研究与发展，49(4)：735-745.

[2] 陈艺云，2022. 基于文本信息的上市中小企业财务困境预测研究[J]. 运筹与管理，31(4)：136-143.

[3] 丁岚，骆品亮，2017. 基于Stacking集成策略的P2P网贷违约风险预警研究[J]. 投资研究，36(4)：41-54.

[4] 段珊珊，朱建明，2016. 基于网络舆情的企业财务危机动态预警[J]. 北京邮电大学学报(社会科学版)，18(6)：31-38，73.
[5] 冯乾彬，向姝婷，蒋为，2023. 数字经济背景下互联网金融企业海外市场进入模式研究——以蚂蚁金服进入“一带一路”沿线国家为例[J]. 技术经济，42(4)：34-54.
[6] 郭全中，李祖岳，2023. 动因、挑战、破局：中国互联网企业出海初探[J]. 新闻爱好者，(5)：15-19.
[7] 蒋殿春，唐浩丹，2021. 数字型跨国并购：特征及驱动力[J]. 财贸经济，42(9)：129-144.
[8] 李美玉，刘洋，王艺璇，等，2023. 基于 Stacking 集成学习的用户付费转化意向预测方法研究——以免费增值游戏为例[J/OL]. 数据分析与知识发现：1-15[2023-03-20]. https：//kns. cnki. net/kcms/detail/10. 1478. g2. 20230317. 1235. 004. html.
[9] 林萍，吕健超，2023. 基于Stacking集成学习的在线健康社区问答信息采纳识别研究[J]. 情报科学，41(2)：135-142.
[10] 刘晓，周荣喜，李玉茹，2023. 基于Stacking算法集成的我国信用债违约预测[J]. 运筹与管理，32(3)：163-170.
[11] 楼润平，李贝，齐晓梅，2019. 中国互联网企业的成长路径、公司战略及管理策略研究[J]. 管理评论，31(12)：11-24.
[12] 吕峻，2014. 基于不同指标类型的公司财务危机征兆和预测比较研究[J]. 山西财经大学学报，36(1)：103-113.
[13] 马述忠，吴鹏，房超，2023. 东道国数据保护是否会抑制中国电商跨境并购[J]. 中国工业经济，(2)：93-111.
[14] 任婷婷，鲁统宇，崔俊，2021. 基于改进 AdaBoost算法的动态不平衡财务预警模型[J]. 数量经济技术经济研究，38(11)：182-197.
[15] 史佳琪，张建华，2019. 基于多模型融合 Stacking集成学习方式的负荷预测方法[J]. 中国电机工程学报，39(14)：4032-4042.
[16] 宋彪，朱建明，李煦，2015. 基于大数据的企业财务预警研究[J]. 中央财经大学学报，(6)：55-64.
[17] 滕晓东，宋国荣，2021. 智能财务决策[M]. 北京：高等教育出版社.
[18] 王静，2020. 我国企业跨国并购的现状、问题及对策建议[J]. 技术经济，39(2)：73-78.
[19] 王言，周绍妮，石凯，2021. 国有企业并购风险预警及其影响因素研究——基于数据挖掘和XGBoost算法的分析[J]. 大连理工大学学报(社会科学版)，42(3)：46-57.
[20] 吴春雷，马林梅，2007. 上市公司最佳资本结构：基于财务预警的实证研究[J]. 经济纵横，(10)：23-25.
[21] 肖毅，熊凯伦，张希，2020. 基于TEI@I方法论的企业财务风险预警模型研究[J]. 管理评论，32(7)：223-235.
[22] 徐继伟，杨云，2018. 集成学习方法：研究综述[J]. 云南大学学报(自然科学版)，40(6)：1082-1092.
[23] 晏艳阳，汤会登，2023. 东道国媒体情绪对中国企业跨境并购的影响研究[J]. 国际贸易问题，(11)：158-174.
[24] 杨勃，齐欣，张宁宁，2020. 新兴市场跨国企业国际化的来源国劣势研究——基于组织身份视角[J]. 经济与管理研究，42(1)：113-125.
[25] 杨剑锋，乔佩蕊，李永梅，等，2019. 机器学习分类问题及算法研究综述[J]. 统计与决策，35(6)：36-40.
[26] 杨连星，2021. 反倾销如何影响了跨国并购[J]. 金融研究，(8)：61-79.
[27] 杨子晖，张平淼，林师涵，2022. 系统性风险与企业财务危机预警——基于前沿机器学习的新视角[J]. 金融研究，(8)：152-170.
[28] 俞锋，池仁勇，2015. 中国企业跨国并购法律风险评价及“浙江模式”总结[J]. 技术经济，34(5)：86-93.
[29] ATHERN K R，DAMINELLI D，FRACASSI C，2015. Lost in translation? The effect of cultural values on mergers around the world[J]. Journal of Financial Economics，117(1)：165-189.
[30] AHMAD M F，LAMBERT T，2019. Collective bargaining and mergers and acquisitions activity around the world[J]. Journal of Banking & Finance，99：21-44.
[31] APPICE A，2015. Lecture notes in Artificial intelligence[C]//CHOWDURY N，CAI X C，LUO C. Boostmf：Boosted matrix factorisation for collaborative ranking. New York：Springer，3-18.
[32] DASARATHY B V，SHEELA B V，1979. A composite classifier system design：Concepts and methodology[J]. Proceedings of the IEEE，67(5)：708-713.
[33] JIA Z H，SHI Y K，YAN C，et al，2020. Bankruptcy prediction with financial systemic risk[J]. The European Journal of Finance，26(7)：666-690.
[34] LAI M T，2022. Analysis of financial risk early warning systems of high-tech enterprises under big data framework[J]. Scientific Programming，2022：9055294.
[35] LUO Y D，2021. New OLI advantages in digital globalization[J]. International Business Review，30(2)：101797.
[36] GEORGIOS S，GEORGIOS P，CONSTANTINE D S，2005. Combining information extraction systems using voting and stacked generalization[J]. Journal of Machine Learning Research，6(3)：1751-1782.
[37] THARWAT A，2016. Principal component analysis a tutorial[J]. International Journal of Applied Pattern Recognition，3(3)：197-240.
[38] TIBSHIRANI R，1996. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society Series B (Statistical Methodology)，58(1)：267-288.

[39] VECCHI A, BRENNAN L, 2022. Two tales of internationalization-Chinese internet firms' expansion into the European market[J]. Journal of Business Research, 152: 106-127.
[40] WOLPPERT D H, 1992. Stacked generalization[J]. Neural Networks, (5): 241-259.

Research on Big Data Early Warning of Financial Risks of Overseas Mergers and Acquisitions of Internet Enterprises Based on Stacking Integrated Learning

Jiang Qiankun, Wang Chengzhe
(School of Economics and Management, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: In the era of digital intelligence, China's Internet enterprises' overseas mergers and acquisitions have risen strongly, but the risks are huge. Machine learning and non-financial information can provide new ideas for such risk early warning. Based on existing research, selecting the big data of 56 overseas mergers and acquisitions of 45 Chinese Internet listed companies from 2013 to 2020, the Stacking integrated learning model was used to mine the big data financial risk early warning factors. The results show that the stacking integrated learning model has better big data early warning effect than other machine learning models. Traditional financial indicators such as operational capacity are still the preferred indicators for big data early warning of financial risks of overseas mergers and acquisitions of Internet enterprises, but innovative non-financial indicators such as stock bar reviews also have important early warning value. The research conclusions provides empirical evidence that Stacking machine learning and stakeholder big data information can help to early warning the financial risks of overseas mergers and acquisitions of Internet enterprises, and provides important reference for Internet enterprises, investors, regulators, etc. to make financial risk control decisions of overseas mergers and acquisitions.

Keywords: Stacking integrated learning; overseas mergers and acquisitions; big data early warning; share bar comments; internet enterprises