实基于XGBoost模型的企业非法集资风险识别

余宗健　汪之鹏　金古阿嘎　陈玥欣　饶师瑗
西南财经大学　四川成都　611130

摘　要：在数字经济时代，利用机器学习方法高效识别企业非法集资风险，愈来愈成为企业数字化监管的重要途径之一。传统的企业风险管理和学术研究多基于静态局部、人工主导等观念，难以应对瞬息万变的企业发展和大数据驱动下的风控需求。本文从动态、客观、机器学习的视角出发，结合法律法规和专家经验，构建企业非法集资的用于识别不同类型关键词词库，基于关键词频率，设计逻辑匹配算法，对企业处罚文本进行识别和标注。在此基础上，建立XGBoost分类模型，应用于企业非法集资风险识别。研究结果表明，与传统机器学习模型相比，XGBoost模型效果最优，实现了企业非法集资风险的精准识别，为企业风控和数字化监管提供参考。

关键词：数字化监管；企业非法集资风险识别；自然语言处理；逻辑匹配算法；XGBoost

Risk identification of illegal fund-raising based on XGBoost model

ZongjianYu, ZhipengWang, AgaJinGu, YueXinChen, ShiyuanRao
Southwestern University of Finance and Economics, Chengdu 611130

Abstract: In the era of digital economy, the use of machine learning methods to efficiently identify the risk of illegal fund-raising has become an important way of digital supervision of enterprises. Traditional enterprise risk management and academic research are mostly based on static local and manual dominated concepts, which are difficult to cope with the rapidly changing enterprise development and the risk control requirements driven by big data. In this paper, from the perspective of dynamic, objective and machine learning, combined with laws and regulations and expert experience, the construction of enterprises illegal fund-raising for the identification of different types of keywords, based on the keyword frequency, the design of logical matching algorithm, enterprise punishment text recognition and annotation. On this basis, the XGBoost classification model is established and applied to the risk identification of illegal fund-raising. The results show that compared with the traditional machine learning model, the XGBoost model has the best effect, realizing the accurate identification of the risk of illegal fund-raising, and providing a reference for the risk control and digital supervision of enterprises.

Keywords: digital supervision; Identification of risk of illegal fund-raising; Natural language processing; Logical matching algorithm; XGBoost

作者简介：
1.余宗健（1983.8—），男，汉族，四川省成都市，工程师，研究生，研究方向：机器学习、金融科技。
2.汪之鹏（1999.9—），男，汉族，安徽省黄山市，本科，研究方向：金融科技、金融工程。
3.金古阿嘎（2000.11—），女，彝族，云南省丽江市，本科，研究方向：计算机应用技术、人工智能。
4.陈玥欣（2002.6—），女，汉族，陕西蓝田，本科，研究方向：自然语言处理、人工智能。
5.饶师瑗（2002.08—），女，汉族，重庆市渝北区，本科，研究方向：人工智能，金融智能。

引言：

近年来，党中央高度重视数字经济，从国家战略层面部署推动其发展。作为一种新兴经济形态，数字经济已然成为中国经济高质量发展的新动能。在此背景下，各行各业纷纷掀起数字化转型的浪潮，运营方式持续迭代，创新业态不断衍生。然而，企业非法集资风险却不断的涌现，例如P2P平台违规乱象，对金融市场和实体经济的稳健运行造成了较大冲击。党的十九大报告提出，要推动互联网、大数据、人工智能和实体经济深度融合，信息技术与企业管理的交叉融合是其中至关重要的一个维度。党的十九届五中全会再次强调，要发展数字经济，推进数字产业化和产业数字化，推动数字经济和实体经济深度融合，打造具有国际竞争力的数字产业集群。国务院日前印发《“十四五”数字经济发展规划》，明确指出“数字经济治理体系需进一步完善”。

新时代诞生新机遇，新时代也催生新挑战，作为实体经济的重要载体，企业在“双循环”的新发展格局中承担着重要角色，是社会运转的细胞，是创造价值的源泉，企业的稳定运营在促进经济增长、技术创新吸纳就业、改善民生等方面具有不可替代的作用。因此，如何在数字经济时代完善企业非法集资风险识别，及时强化有效预防和监管，保障企业健康，成为实现企业未来可持续发展、促进实体经济良性增长的重要课题。

本研究在收集企业处罚文本数据的基础上，企业非法集资风险的关键词识别词库，设计逻辑匹配算法，对企业处罚文本进行标注，筛选出涉及非法集资风险的相关企业。在此基础上，提取相关企业基本面数据，并采用逻辑回归、决策树、XGBoost等模型进行比较实验，验证模型对企业非法集资风险的识别效果。最终，通过研究为企业科学风控和政府有效监管提供参考和支撑。

一、文献综述

1.机器学习应用于企业风险的分析

传统的企业风险管理多基于简单通俗的方法，如调查访谈、报表审查、建立信息系统等，随着企业风险的逐渐复杂化、管理的日益技术化，学术界也开始运用机器学习和深度学习等智能算法来研究企业风险。

近年来，国内学者利用机器学习方法对企业风险展开了大量分析。张大斌等（2015）设计差分算法并建立聚类模型评估了中国上市公司的信用度，并与遗传算法、决策树模型等进行了比较。方匡南等（2016）指出运用经典的逻辑回归来建立企业运营风险预警模型的效果欠佳，进而提出了基于网络关联结构的改进逻辑回归模型。胡贤德等（2017）参考群智能萤火虫算法，提出改进的离散型萤火虫算法，并将其引入BP神经网络，用于中小微企业的运营风险评估。熊正得等（2018）借助因子分析法，基于A股上市企业的财务数据构建了风险评价体系，并应用逻辑回归方法对不同组样本的违约概率进行了测度。沈彦菁等（2019）基于企业征信系统中三万余家小微企业的财务数据，运用支持向量机模型，初步找到了几个可以帮助小微企业增信融资的重要指标。

可见，机器学习方法在企业风险分析领域的应用已较为广泛成熟，且衍生出了信用评估、财务造假等多个细分枝和创新视角。因此，本研究针对企业非法集资风险，综合运用逻辑回归、决策树、XGBoost等各类模型算法，增强对企业非法集资风险的识别能力。

二、研究设计

1.研究数据来源

本研究通过中国工商行政管理总局、国家市场监督管理总局等公开平台收集企业数据，获得5万余家企业的法律处罚记录、经营状态、从业人数等30维企业的基本特征数据，时间跨度为2010年1月1日至2021年12月31日。

2.面向企业非法集资风险的关键词词库构建

本研究基于获取的企业法律处罚记录（每家企业对应唯一的法律处罚记录）。通过研究法律处罚记录文本的特点，同时参考《中华人民共和国公司法》、《中国人民共和国刑法》、《企业经营异常名录管理暂行办法》、《非法金融机构和非法金融业务活动取缔办法》等各类法律条文。本研究挖掘构建面向企业非法集资风险的关键词词库，用以匹配识别不同类型。面向企业非法集资风险的关键词词库，如下表1所示。

需要指出，上述关键词词库中的词语顺序并不是任意的，而是综合考虑其在正式法律条文和实际处罚文本中的出现频率和相关程度，根据该关键词对于类型识别的重要性和特异性先后排序得到。

3.识别匹配算法设计

基于企业非法集资风险的关键词词库构建，本研究可以根据法律处罚记录文本对获取企业进行非法集资风险标记。然而，如何标记需要一定的识别匹配算法来实现。本研究提出了一种识别匹配算法，具体而言，本研究将企业的法律处罚记录文本与面向企业非法集资风险的关键词词库进行匹配，将出现在法律处罚记录文本中的关键词在词库中的倒序排名相加，再除以从1到该词库长度（关键词个数）的累加和，最后放大100倍，作

表1　面向企业非法集资风险的关键词词库

风险类型	典型法律处罚记录文本示例	关键词词库
非法集资风险	例如，房地产广告中出现融资或者变相融资的内容，含有升值或者投资回报的承诺。	非法集资、非法融资、非法吸收公众存款、变相吸收公众存款、非法获取资金、集资诈骗、集资、融资、集资款、吸收公众存款、吸收资金、承诺高额回报、承诺高额固定收益、高额利息、高额回报、高额固定收益、非法占有、返本销售、售后包租、约定回购、编造虚假项目、订立陷阱合同、虚假转让股权、发售虚构债券、假借境外基金、发售虚构基金、假冒保险公司、伪造保险单据、肆意挥霍集资款、携带集资款逃匿、抽逃资金、转移资金、隐匿财产、逃避返还资金、隐匿账目、销毁账目、假破产、假倒闭、数额较大、数额巨大

为匹配得分。最终，本研究根据提出的识别匹配算法，对所有受过法律处罚的企业进行匹配识别。经过一系列处理，成功匹配具有企业非法集资风险的企业共有23723家。

三、面向企业非法集资风险的模型构建

1.数据预处理

根据企业的法律处罚记录文本，本研究利用识别匹配算法给企业添加了非法集资风险标签。在此基础上，本研究利用收集的企业特征数据建立机器学习模型，对企业非法集资风险进行建模识别。首先，本研究采用较为常见的做法进行缺失值填充，即数值型特征用均值填充、类别型数据用众数填充，另外，本研究对类别型特征进行了One-hot编码，从而特征数据全部转化为了数值型。

2.分析模型构建

本研究经过数据特征探索后，采用收集的30维基本面指标作为数据特征，以AUC值（Area Under Curve，即，ROC曲线下与坐标轴围成的面积）为模型性能的衡量标准，建立XGBoost分类模型。同时，为了考量模型表现，本研究构建了其他类模型进行对比实验，对比模型包括逻辑回归、决策树、随机森林、GBDT、LightGBM。

3.实验结果

本研究将所有企业样本按7：3的比例，随机划分为训练集和测试集，依次建立了逻辑回归、决策树、随机森林、GBDT、LightGBM、XGBoost模型，分别应用于企业非法集资风险识别，并在模型分析过程中通过剪枝（决策树）、调参、交叉验证等手段来改进优化模型预测的效果。

为了直观比较不同模型之间的效果差异，本研究绘制不同模型应用于企业非法集资风险识别的ROC曲线，如图1所示。实验结果可以看出，基础的逻辑回归模型表现相对较差，而随机森林、GBDT、LightGBM、XGBoost表现相对更优。其中，XGBoost又相对更具优势，AUC值达到了0.89。

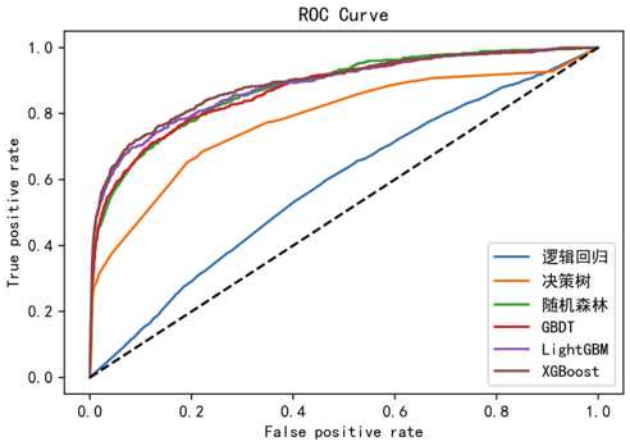

图1　不同模型应用于放贷风险的ROC曲线

4.实证分析

在实验结果的基础上，本研究设计了实证分析实验，对暂无法律处罚记录的企业进行分析，验证XGBoost模型在企业非法集资风险识别领域的泛化效果。在实证分析实验中，本研究重新收集了5000家无处罚记录的企业及其相应特征数据为验证集（人工对企业非法集资风险进行评估），应用训练好的XGBoost模型预测出了企业存在非法集资风险的概率。根据实证结果，本研究对存在非法集资风险进行人工调查，存在非法集资违法犯罪行为的企业识别准确率达到78%，证明本研究提出的XGBoost模型对企业非法集资风险识别具有非常优秀的效果和泛化应用性。

四、总结与展望

数字经济给各行各业创造了新机遇，也给实体经济的稳定运行带来了新挑战，持续高效进行企业风险管理的重要性不言而喻，而面对当前复杂多变的企业风控现状，传统的手段已显乏力，引入前沿的模型算法越来越凸显出必要性和优越性。本研究紧跟前沿，基于机器学习的视角，对企业风险识别问题展开了深入探究，不

仅基于企业的法律处罚记录文本信息自主构建了非法集资风险的关键词词库，设计了逻辑匹配算法，还利用企业的净利润等30维基本面特征，综合建立了逻辑回归、XGBoost等二分类模型，并通过交叉验证等多种方式不断优化效果，ROC曲线表明模型表现良好。

可以看出，本研究成果具有一定的现实意义，能够为企业风控和监管提供参考。首先，本研究成果可实际用于企业潜在非法集资风险识别，只需模型对应的特征数据输入，便可预测暂未受法律处罚或无处罚记录的企业存在非法集资风险的概率，从而，政府监管部门可在海量企业中实现分级管理和靶向跟踪，助力监管的方向性。其次，利用机器学习方法探查企业风险是当前学术研究的热点之一，本研究的思路和方法或许能为这方面的进一步研究提供一些启发。

参考文献：

[1]方匡南，范新妍，马双鸽.基于网络结构Logistic模型的企业信用风险预警[J].统计研究，2016，33（04）：50-55.

[2]胡贤德，曹蓉，李敬明，阮素梅，方贤.小微企业信用风险评估的IDGSO-BP集成模型构建研究[J].运筹与管理，2017，26（04）：132-139+148.

[3]沈彦菁，张榕薇，朱维聪.基于机器学习方法的小微企业融资特征分析研究—关于嘉兴市33305家小微企业的SVM模型实证分析[J].金融经济，2019，（10）：105-108.

[4]熊正德，张帆，熊一鹏.引入WFCM算法能提高信用违约测度模型准确率吗?—以沪深A股制造业上市公司为样本的实证研究[J].财经理论与实践，2018，39（01）：147-153.

[5]张大斌，周志刚，许职，李延晖.基于差分进化自动聚类的信用风险评价模型研究[J].中国管理科学，2015，23（04）：39-45.