基于随机森林的第三方支付违规风险预警研究

方若男,骆品亮
(复旦大学 管理学院,上海 200433)

摘 要:立足于第三方支付行业发展中存在的突出问题,本文给出基于随机森林的违规风险预警机制并讨论具体实施。首先,构建风险预警指标体系,结合机器学习中的随机森林算法,提出风险预警机制。然后,以已获支付牌照的271家企业为样本,验证所提出的违规风险预警机制的有效性。通过对比随机森林模型和Logistic模型的判定结果,发现随机森林显著降低了一类错误率和二类错误率,模型正确预测率高达99.01%。最后,通过对指标体系中的重要变量进行分析,提出具体应用措施及相应的风险监管建议。

关键词:第三方支付牌照;风险预警;随机森林;风险监管
中图分类号:F062.9 **文献标志码**:A **文章编号**:1002—980X(2020)9—0011—11

作为互联网时代支付产业的创新形式,第三方支付在我国金融服务业中扮演着重要的角色,对实体经济的转型发展也发挥着越来越重要的作用。尽管监管趋严,但第三方支付企业的各种违规行为屡禁不止,对支付产业生态产生严重的负面影响。随着市场准入限制的进一步放开①,如何从根源上进行风险预警显得尤为重要。因此,建立完善的第三方支付违规风险预警机制,构建科学的风险评估指标体系和评估方法,具有重要的理论和现实意义。

目前关于第三方支付风险管控的相关研究主要集中在3个方面。其一,第三方支付风险的构成。杨彪[1]总结了第三方支付的4类风险,包括金融系统性风险、信息系统与操作风险、市场环境风险和市场退出潜在风险,其中市场环境风险的具体表现为恶性竞争、洗钱、套现、赌博等形式,市场退出潜在风险包括未取得行政许可的机构退出、取得了行政许可但因经营不善而退出两大类。胡娟[2]指出了第三方支付风险的具体构成,包括巨额沉淀资金风险、洗钱风险、信用卡套现风险、消费者权益保护风险等法律风险,业务设施技术风险以及跨境支付风险。Yang等[3]从用户感知和接受度出发,指出支付平台与用户之间存在着信息不对称风险,尤其是绩效风险、财务风险和隐私风险会显著影响支付平台的接受度。其二,风险监管机制设计。针对互联网金融的监管难点,修永春[4]提出了三元监管模式,并分析其分阶段运行机制。Xu等[5]指出中国互联网金融市场面临的监管挑战,对目前存在的"自上而下"监管方式的有效性展开探讨。针对如何防范风险,学者们从监管体制、市场准入和退出机制等方面提出了许多建议[6-7]。其三,风险预警。关于信用风险的影响因素方面,相关研究发现公司的发展前景、盈利能力、公司规模、财务杠杆等因素会影响企业的财务欺诈和信息披露违规[8-11],投资者关系管理和内部控制质量对企业未来的违规倾向有显著影响[12]。在信用风险预测方面,Logistic模型是最常用的风险计量模型[13-14]。最近几年,数据挖掘和机器学习等新兴工具在风险预警中的应用日受重视。涂艳和王翔宇[15]通过对比试验研究,发现运用机器学习来预测借款人违约行为的准确率远高于传统的回归模型,特别是随机森林算法的效果尤佳。林成德和彭国兰[16]发现,通过随机森林方法得到的指标体系能更有效地反映企业的信用状况。目前,随机森林算法被广泛应用各个行业中企业信用风险的评估中[17-20],而支持向量机等其他机器学习方法也受到越来越多的关注[21]。

总体而言,目前针对第三方支付违规风险的研究方兴未艾,系统深入的定量研究亟待开发。对于随机森林算法在风险预测模型中的应用,既有研究大多只是直接将其应用于某个具体行业,并未与传统模型方法进行充分的对比分析,也缺乏对指标体系的详细分析。在此背景下,本文将随机森林模型方法应用于第三方支付违规风险的预警研究中,构建预警指标体系,给出预警机制及其具体实施方案。

收稿日期:2019—12—12

作者简介:方若男(1993—),女,河南三门峡人,复旦大学管理学院博士研究生,研究方向:网络竞争与规制、平台经济;骆品亮(1969—),男,福建惠安人,博士,复旦大学管理学院教授,博士研究生导师,研究方向:产业组织理论、网络竞争与规制、创新动力学。

① 2018年博鳌亚洲论坛上,国家主席习近平在其主旨演讲中宣布中国将大幅放宽包括金融业在内的市场准入,中国人民银行行长易纲随后表示已经放开银行卡清算机构和非银行支付机构的市场准入限制。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

本文首先基于经典风险理论并结合第三方支付行业的特殊性，从企业基本特征和经营行为两个方面出发，用企业自身特征、企业发展特征、创新行为、投融行为、信用行为5个维度的变量构建了风险预警指标体系。在此基础上，运用随机森林模型提出了违规风险预警机制。然后，将样本随机划分为训练集和测试集，并运用十折交叉验证（10-fold cross-validation）评估随机森林模型的准确度。通过对比Logistic回归模型与随机森林模型的结果，验证了所提出的风险预警机制的有效性。最后，根据模型预测效果及对变量的重要性分析，给出相应的监管建议。

与既有研究相比，本文的主要区别体现在两个方面：一是研究对象的不同。本文选取目前备受关注的第三方支付企业进行研究，由于支付牌照存在的特殊性以及行业本身存在的“互联网+”特征，对牌照申请企业的资质和信用风险进行正确合理的审视至关重要。二是研究方法的差异。通过将定性研究与机器学习算法相结合，本文提出了违规风险预警机制，克服了传统模型的局限性，提高了风险预警的有效性。同时，本文在所提出的预警模型基础上，进一步提炼出重要变量并展开分析，对如何把控市场准入机制、如何实施监管进行了详细描述和研究。

一、风险预警机制设计

构建第三方支付的风险预警指标体系，根据随机森林算法提出第三方支付违规风险预警机制，以271家持牌支付企业为样本检验基于随机森林的风险预警机制的有效性。

（一）风险预警指标体系构建

根据国内外相关研究，企业特征对企业的违规风险有重要的影响。例如，企业的发展前景、盈利能力、公司规模、财务杠杆等会对其财务欺诈问题和信息披露违规问题产生重要影响[6-9]；而企业违约信号、投资者关系管理和内部控制质量也会显著影响企业未来的违规倾向[10]。需要指出的是，对于第三方支付行业，其独特的行业特性蕴含着相应的风险特征：一方面，第三方支付通过发放支付牌照的方式来认定企业的经营资格，企业自身特征（如类型、规模、媒体曝光度等）展示了企业的实力，但也暗示了相应的风险；另一方面，第三方支付行业具有显著的“互联网+”特征，在基础业务的基础上还有各种增值业务，而且大部分第三方支付企业都有技术背景，且备受资本热捧②。因此，在分析第三方支付企业的违规风险时，应特别关注其经营行为，特别是企业创新、投融资及信用行为。

通过对研究样本中的18家违规企业的违规行为进行跟踪分析，本文发现第三方支付企业的违规行为主要来自于两方面：一是企业自身条件未能满足相关管理规定，或是业务不符合要求，或是在财务、数据等方面有违法行为；二是企业为了自身利益进行挪用备付金、私自转让牌照等违规操作。

综上所述，本文从企业特征和经营行为两个方面构建第三方支付的风险预警指标体系，见表1，其中企业特征包括自身特征和发展特征，经营行为包括创新行为、投融行为和信用行为。

表1　风险预警指标体系变量汇总

解释变量来源	解释变量类别	解释变量名称	变量描述与测量
企业特征	自身特征	企业类型	分类变量：分为“有限责任公司”与“股份有限公司”两类
		企业规模	分类变量：按企业注册资本金额分为4类。小于5千万；5千万至1亿3千万；1亿3千万至8亿；大于8亿
	发展特征	获牌时间	从注册成立至拿到牌照的天数
		媒体报道	关于该企业的媒体报道数
		投资方来源	分类变量：无任何背景为0；有上市背景为1；有国资背景为2；同时有上市和国资背景为3
经营行为	创新行为	资质证书	企业所获的资质证书数
		注册商标	企业所注册商标数
		所获专利	企业获得专利数
		软件著作	企业拥有软件著作数
	投融行为	对外投资	企业对外投资数
		融资历史	企业融资次数
	信用行为	工商变更	企业工商变更次数
		失信人信息	企业出现失信人信息的次数
		经营异常	企业出现经营异常的次数
		行政处罚	企业受到行政处罚的次数
		股权质押	企业股权质押次数

1. 企业特征

企业自身特征包括企业类型和企业规模两个变量。关于企业类型，本

② 根据零壹财经·零壹智库发布的《2017全球金融科技发展指数（GFI）与投融资年报》，2017年支付行业获得93笔融资，总额约265亿元，是金融科技中4个融资金额达到百亿元的领域之一（另外3个是借贷、汽车金融和互联网保险）。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

文将其分为“有限责任公司”和“股份有限公司”两种，由于有限责任公司和股份有限公司在股东数量、注册资本和组织机构权限等方面有诸多不同，因而，本文认为其将对企业行为产生影响。关于企业规模的重要性已引起国内外研究的关注，本文按企业注册资本金额将其分为4类：小于5千万、5千万至1亿3千万、1亿3千万至8亿和大于8亿。

企业发展特征包括获牌时间、媒体报道和投资方来源3个变量。其中，获牌时间系指企业从注册成立至拿到牌照的时间。关于媒体报道对企业在资本市场表现的影响，已有研究发现媒体报道可以通过媒体信息治理和声誉治理等多种路径影响公司行为，比如，媒体监督会提高被报道企业的公众关注度和企业透明度，产生聚光灯效应，影响企业行为[22]；越受媒体关注的企业，越乐意改善社会责任表现[23-24]；媒体关注度不仅直接影响企业社会责任的履行，还会通过对其他利益相关者的影响给企业实施压力，促使企业的自律行为[25]；一些学者认为媒体作为外部监督机制能够发挥积极作用[26-29]，而游家兴和吴静[30]研究发现新闻报道所传递出的媒体情绪会传递负面影响。因此，本文也将其作为模型输入自变量，分析其对公司行为的影响。此外，根据权小锋等[12]对于投资者关系管理的研究结果，本文加入投资方来源变量，并根据企业背景将其分为4类：无任何背景、有上市公司背景、有国资背景、同时有上市和国资背景。

2. 经营行为

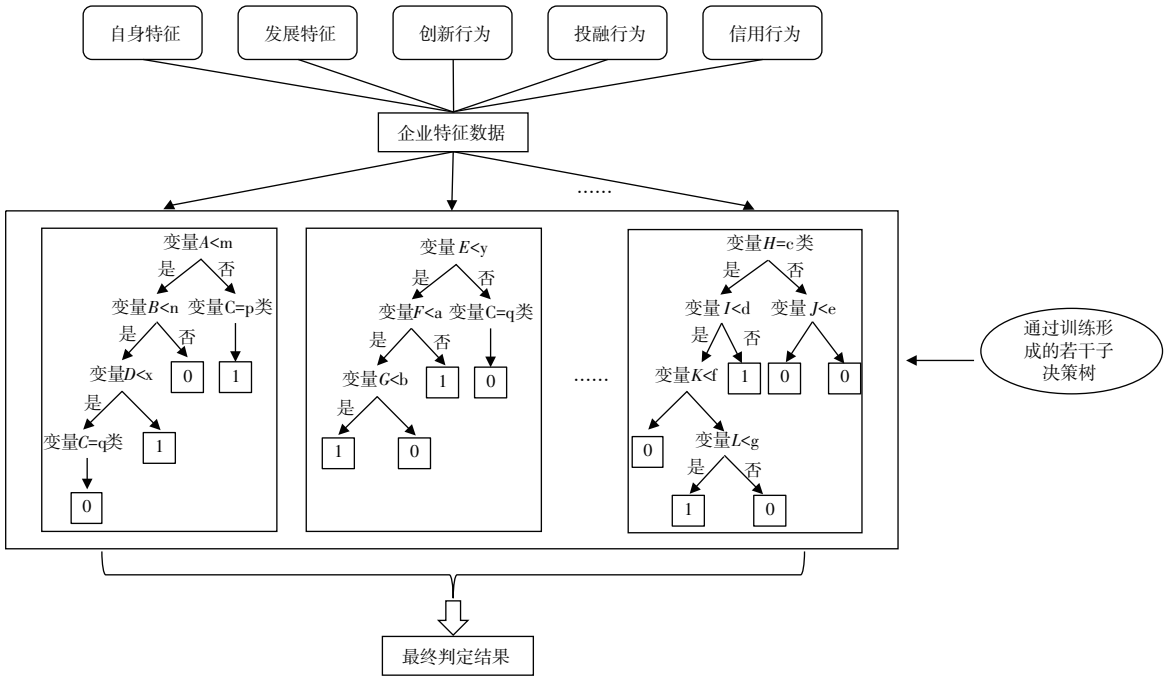
创新行为包括资质证书、注册商标、所获专利和软件著作4个变量，这些变量反映了企业在知识产权、研发、创新方面的投入与成果，是企业经营实力和竞争力的体现。

投融行为包括对外投资和融资历史两个变量。其中，对外投资的目的是获得投资收益，但也有可能遭受经济损失。企业融资则从某种角度反映了企业信息披露违规的可能性。实际上，当有较好的发展前景时，公司进入资本市场融资的动力更强，此时公司的管理层往往不愿因犯错而失去宝贵的融资机会，因此公司违规披露信息的可能性就比较低[10]。

信用行为包括工商变更、失信人信息、经营异常、行政处罚和股权质押5个变量，这些变量分别反映了企业的经营风险和财务风险。

（二）随机森林预警过程

根据Breiman[31]提出的随机森林算法框架，本文构建了如图1所示的基于随机森林的违规风险预警机制。



注：$A \sim K$表示文中所构建指标体系中的变量，a～y表示变量的值。如“变量A<m”可能表示“媒体报道<300条”。此处描述的正是预警机制的构建过程，由于无法预知各个变量的重要性及判别标准，通过随机森林模型对数据集进行训练形成子决策树后，进而可对新企业的风险进行判定

图1　基于随机森林的风险预警过程示意图

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

风险预警的具体过程如下③：

首先，对于每一家企业，运用所构建的预警指标体系变量刻画其特征x，作为随机森林模型的输入变量，因变量y为0-1二值变量，分别代表企业有严重违规和无严重违规。

然后，从全集数据中随机选取若干变量和若干样本数据，将其划分为n个子集，构成n个训练集，以此来增加各个子分类模型的差异。

最后，随机森林便对训练集中的数据开始“学习”过程。基于每个训练集中的数据和变量，随机森林模型对样本数据信息不断进行提炼，通过递归的方式计算得到最优的划分特征以及特征的划分节点，依次进行根节点至叶节点的变量特征选择，构建出若干子决策树，这些子决策树共同构成训练模型。

当需要对新的样本数据进行判定时，将该企业的特征变量输入模型，每一棵子决策树都能够给出一个判定结果，模型基于多数原则输出最终的判定结果：0或1。

（三）预警机制有效性检验

1. 数据来源及变量描述性统计分析

选取中国人民银行官方网站公布的《支付业务许可证》获取名单中的271家企业为研究对象，通过天眼查、启信宝收集到这些企业的相关信息进行研究。这271家企业的牌照类型覆盖了央行所规定的三大类业务类型，可以代表整个第三方支付市场。同时，天眼查和启信宝的数据来源于全国企业信用信息公示系统、中国法院裁判文书网、中国执行信息公开网等100多家官方网站，因此，研究对象的选择具有代表性和普适性。

研究样本分为两类：严重违规（被注销牌照）、无严重违规。在被注销牌照的33家企业中，将因为合并注销的11家企业和主动申请注销的4家企业重新划分为无严重违规类别。核心被解释变量“企业违规情况”为虚拟变量：若企业发生严重违规致被吊销牌照取1，否则取0，见表2。核心解释变量即表1中所构建的风险预警指标体系变量。

表2 研究样本牌照情况

违规类别	牌照情况	样本量	占比	因变量取值
严重违规	因严重违规被注销牌照	18	6.64%	1
无严重违规	因业务合并被注销牌照	11	4.06%	0
	主动注销牌照	4	1.48%	0
	未注销牌照	238	87.82%	0

注：数据来源为根据央行公布数据整理。

所有变量的描述性统计分析见表3。可以看到，被解释变量的均值非常接近0，这是因为原始数据中“有严重违规”的企业数量比起“无严重违规”太少，因此，本文在后面的分析中对数据先做了均衡处理。在企业特征变量中，从企业类型变量看，有限责任公司占多数。从获牌时间和媒体报道看，样本中所有企业这两个变量差异非常大。在经营行为变量中，不同企业注册商标数差别非常大，其余的创新行为特征变量差异相对小很多。从信用行为看，绝大多数企业均具有良好的信用资质，体现为工商变更、失信人信息、经营异常、行政处罚和股权质押变量均值都比较小。

表3 变量的描述性统计分析

类别	变量	观测个数	平均值	标准差	最小值	最大值
被解释变量	企业违规情况	271	0.0664	0.2494	0	1
企业特征	企业类型	271	0.9446	0.2290	0	1
	企业规模	271	2.0480	0.7362	0	4
	获牌时间	271	1580.4000	1293.6990	127	6849
	媒体报道	271	163.700	311.9082	0	3441
	投资方来源	271	0.8229	1.0462	0	3
经营行为	资质证书	271	1.4650	1.7250	0	10
	注册商标	271	31.5000	101.7880	0	1440
	所获专利	271	3.6200	13.7348	0	138
	软件著作	271	13.1200	24.0962	0	253
	对外投资	271	2.0520	3.8168	0	34
	融资历史	271	0.4170	0.9811	0	8
	工商变更	271	18.9300	20.8510	0	242
	失信人信息	271	0.1587	1.2325	0	15
	经营异常	271	0.1993	0.5137	0	3
	行政处罚	271	0.2214	0.5190	0	3
	股权质押	271	0.8155	2.7734	0	31

注：数据来源为计算整理而得。

2. 机制有效性检验

基于R语言实现Logistic回归模型和随机森林模型的构建④。

首先是对所构建指标体系的有效性进行检验。在随机森林模型中，分别选取1个、2个、4个、8个和16个变量放入模型，见表4，发现当变量个数为16个时模型整体的误差最小，只有0.010。

③ 在Breiman[31]关于随机森林算法框架的基础上，本文构建了风险预警的指标体系，并实际运用了该算法框架，形成风险预警过程。

④ 由于原始数据中出现严重违规以致被吊销牌照的数量比例较小，在随机森林模型中首先使用SMOTE算法对数据进行预处理，实现不平衡数据的均衡化。SMOTE算法处理后的数据中“无严重违规”和“严重违规”分别有360个和378个。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

同理，在进行Logistic回归时，分别用4种不同的变量组合构建模型，来检验不同类别信息的预测能力。模型1基于企业特征进行判别；模型2基于企业自身特征和信用行为；模型3基于企业自身特征、发展特征和信用行为；模型4则包括所有5类信息变量。判别模型的相关拟合优度指标见表5，从表5可以看到，模型4的准确率和AUC（area under ROC curve）值[⑤]远高于其他3个模型，赤池信息准则（Akaike information criterion，AIC）值小于其他3个模型，这表明包含所有变量的模型拟合效果最优，进一步表明应该将所有变量都放入模型中。

在初步验证了所构建的风险预警指标体系的有效性后，就可以对模型预测度进行比较。在评估模型准确度时，本文通过十折交叉验证验证算法的准确性。具体而言，将数据集分成10份，轮流将其中9份作为训练数据，1份作为测试数据进行试验，最后将10次结果正确率的平均值作为对算法精度的估计。

随机森林与Logistic模型预测对比结果见表6，由表6可以看出：随机森林模型的正确预测率远远高于Logistic模型，Logistic模型正确预测率只有80.33%，而随机森林达到了99.01%，表明随机森林模型对于评判相关企业是否会出现严重违规的准确度高达99.01%。

除此之外，随机森林模型的一类错误率和二类错误率都远低于Logistic回归，模型敏感度达到了100%，这意味着所有有严重违规行为的第三方支付企业都能被正确识别。模型精确度为98.02%，这表示在被模型预测为有严重违规行为的企业中，实际上确实有严重违规行为的企业比例为98.02%。通过验证模型性能常用的指标AUC值来看，随机森林模型的AUC值远远高于Logistic模型，也进一步验证了随机森林模型预测精度更高。

表4　随机森林模型中变量选择的初步结果

选取变量个数	16	8	4	2	1
算法误差	0.010	0.012	0.016	0.051	0.096

注：数据来源为计算整理而得。

表5　不同变量组合下Logistic回归模型拟合效果

判别指标	模型1	模型2	模型3	模型4
赤池信息准则（AIC）	110.230	92.571	98.440	64.401
残差（residual deviance）	90.230	72.571	68.440	22.401
AUC指标	—	0.4706	0.4712	0.6896
准确率	50%	48%	49%	80.33%

注：数据来源为笔者计算整理而得。

表6　随机森林模型与Logistic模型结果对比

模型	正确预测率	一类错误率	二类错误率	AUC指标
Logistic模型	80.33%	6%	33.33%	0.6896
随机森林模型	99.01%	1.98%	0	0.9903

注：数据来源为笔者计算整理而得。

二、风险预警机制应用分析

在风险预警机制的具体实施中，可以先通过关键指标进行初步评估，筛选出风险较高的企业并预警。对通过关键指标筛选标准的企业，则进一步结合完整的指标体系和模型进行风险评估。本节通过对指标体系中的重要变量的分析，并结合典型案例，阐述了预警机制的具体应用，并提出相应的监管政策建议。

（一）重要变量分析

根据平均精度降低度（mean decrease accuracy，MDA）和平均基尼指数降低度（mean decrease gini，MDG）两个指标可以识别重要变量。MDA描述的是一个变量被随机数替代后模型结果准确度的降低程度，值越大表示该变量的重要性越高；MDG描述的是一个变量对分类树上每个节点分类效果的贡献，值越大表示该变量的分类效果越好。运算结果如图2所示，通过综合对比分析，本文发现媒体报道、注册商标数以及获牌时间是最重要的3个变量。

⑤ AUC值表示ROC（receiver operating characteristic）曲线下与坐标轴围成的面积，ROC曲线的纵轴代表“真阳性率”，横轴代表“假阳性率”。AUC值越大，对应的分类器效果越好。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

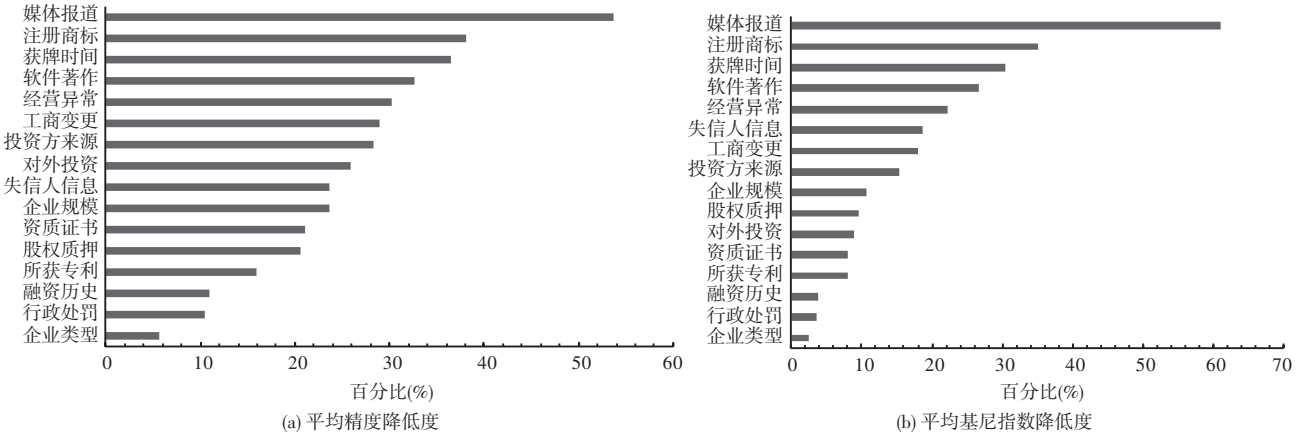
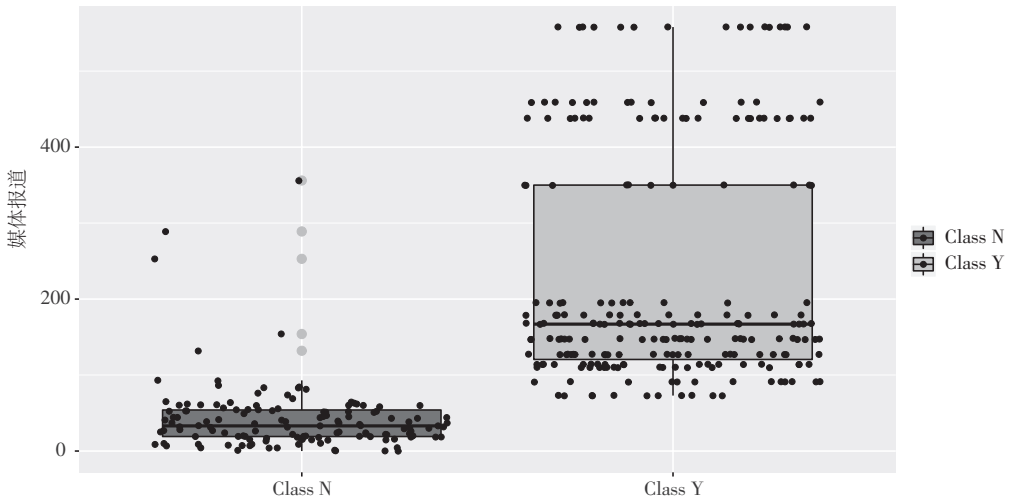
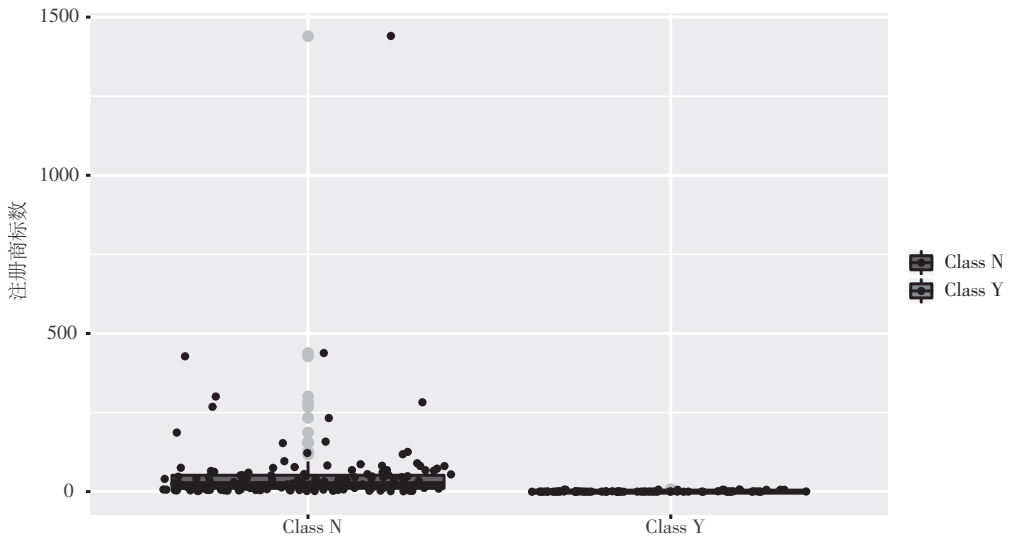


图 2　指标体系中所有特征重要性排序

进一步,给出媒体报道、注册商标数以及获牌时间这 3 个变量在两个类别(严重违规、无严重违规)中的分布,如图 3 所示。需要特别提醒的是,媒体报道、注册商标数这两个变量目前恰恰是央行所忽略的。以下对各变量的重要性逐一进行分析。

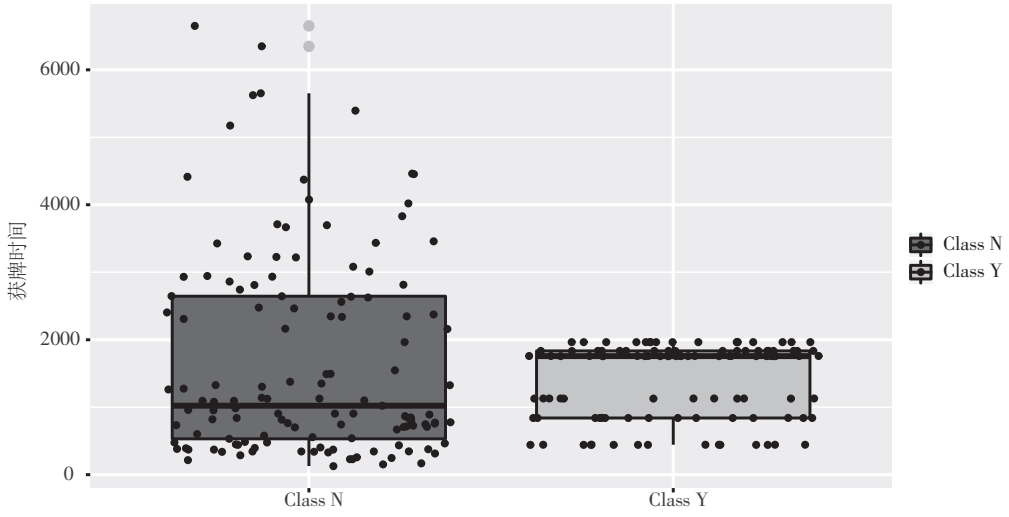


(a) 媒体报道



(b) 注册商标数

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net



（c）获牌时间

Class N为“无严重违规”类别的企业；Class Y为“严重违规”类别的企业

图3　媒体报道、注册商标数和获牌时间3个变量类别分布图

关于获牌时间，该变量代表企业从注册成立到取得牌照的时间。可以看出，被注销牌照的企业的获牌时间变量值普遍较低。企业在发展初期容易出现过度投资、盲目扩张的问题，Cefis和Marsili[32]指出企业生存机会随着企业年龄和成长而增加。肖兴志等[33]通过分析494家企业在6年观测期的表现，发现有28.7%的企业选择退出，平均生存年数仅为3.7年，而且战略性新兴产业的退出风险更大。事实上，该变量的重要性在央行目前的《支付业务许可证》申请要求里已有所体现：央行规定外资企业截至申请日，应连续盈利2年以上。不过，获牌时间变量虽然对整个风险预警机制的贡献很大，但是从图3(c)可以看到，只凭借这一个变量并不能达到很好的分类效果，在判别时仍需结合预警机制中的其他变量。

关于媒体报道对企业运营的影响，一方面，媒体报道数量反映了企业的媒体关注度，这会对公司管理层产生压力效应[34-35]；另一方面，频繁在媒体曝光的公司更容易受到大众的关注，知名度会影响到企业从事违规活动的可能性。此外，已有许多研究指出，新闻报道所传递出的媒体情绪会影响到企业行为。作为信息中介，媒体搜集、整理并传播相关信息，不仅拓宽了企业的外部信息渠道，也加深了市场各方对公司的了解。本文中媒体报道变量在两个类别中的分布如图3(a)所示。与既往研究不同，本文发现违规企业的媒体报道变量分布在110～370次。换言之，媒体报道数低于110次或者高于370次都传递了“企业不会违规”的正信号。通过对样本中企业的新闻报道进一步细分为“正面”和“负面”报道后，本文发现，对于新闻报道较多（新闻报道数高于370次）的企业，其报道内容中正面与负面比例相当，或者正面报道所占比例远高于负面报道比例；而对于新闻报道数相对较低的企业，说明企业本身在媒体中的关注度就很低，相关的新闻报道几乎都是各个媒体转发的央行处罚新闻。从实际情况来看，大部分第三方支付企业规模并不大，在大众和媒体中的知名度都不高。除了支付宝、银联、拉卡拉等少数大型企业，大多数第三方支付机构的规模都很小，在新闻媒体中的活跃度很低，因此出现媒体报道时以负面新闻报道居多，所以媒体报道数低反而表明其负面新闻少，是一种正信号。如支付宝、银联等的大型第三方支付企业，这些企业在媒体上相对活跃许多，媒体报道数都非常高，佐证了媒体作为外部监督机制能够发挥积极作用。

关于企业拥有的注册商标数的重要性，从理论上讲，商标不仅是企业知识产权的组成部分，也是一项重要的无形资产，反映了企业技术创新的实力，它承载着企业的声誉，帮助企业实现市场竞争优势[36]。实际上，随着第三方支付产业竞争的加剧，支付企业呈现多样化，包括银联系、电商系、社交系、移动运营商及通信设备制造商、独立第三方等，运营模式也不断创新，形成“支付+”生态体系。在第三方支付产业链竞争中，商标所传递的信号作用越来越重要。从实际数据来看，合规企业的注册商标数远远高于违规企业，发生严重违规被注销牌照的企业中，注册商标数几乎都为0或者只有个位数。本文的研究结果进一步佐证了潘镇和鲁明泓[37]提出的注册商标的等级对中小企业的业绩有着重要影响的结论。

重要性紧排在这3个变量之后的是软件著作和经营异常两个变量。软件著作权作为知识产权的一部

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

分，是企业重要的无形资产。一方面，软件著作权代表了企业的科技实力，能够提升企业信誉；另一方面，作为知识产权，拥有软件著作权的企业能够享受到国家的各种优惠政策，对其长期发展及可持续经营都有积极作用。更重要的是，对于互联网企业来说，软件著作权能在一定程度上反映其业务的丰富性及多样性，可以提高用户的支付体验。

是否有经营异常则代表了企业的征信情况，体现了企业的风险信息。此外，工商变更次数的重要性也不容小觑，尤其是股东变更。李维安和李建标[38]发现控股股东变动的次数与企业信用呈反向变动关系：股东变动的次数越多，企业信用下降的程度会越大。本文关于第三方支付行业的研究发现支持其结论。

关于企业注册资本的重要性，虽然央行在《支付业务许可证》申请要求里有“企业应有符合规定的注册资本最低限额”的相关规定，但是实际上对企业违规经营的影响并不大，本文认为央行日后可适度调整该项规定。重点关注上文所提到的几个变量，在此基础上结合企业注册资本等变量来综合分析。

（二）典型案例分析

关于媒体报道、注册商标数以及获牌时间3个变量重要性的实际案例代表见表7，其中SXLH公司、CSSL公司和BJJG公司为负面案例代表，YLSW公司、GZYL公司和JSGX公司为正面案例代表⑥。

表7 典型案例重要变量及其他特征变量对比

企业		SXLH	CSSL	BJJG	YLSW	GZYL	JSGX
重要甄别变量	媒体报道（条）	137	167	168	2228	418	45
	注册商标（个）	6	0	0	428	49	23
	获牌时间（天）	1650	1288	1128	3402	3428	3170
其他特征变量	软件著作（个）	0	8	0	220	33	26
	经营异常（次）	0	0	1	1	0	0
	工商变更（次）	4	2	13	40	11	5
	行政处罚（次）	0	1	1	0	0	0
	企业类型	有限责任公司	有限责任公司	有限责任公司	股份有限公司	有限责任公司	有限责任公司
	企业规模（元）	5千万至1亿3千万	小于5千万	小于5千万	1亿3千万至8亿	1亿3千万至8亿	5千万至1亿3千万
	投资方来源	上市背景+国资背景	无任何背景	国资背景	国资背景	国资背景	国资背景
	资质证书（个）	0	1	0	1	5	4
	所获专利（个）	0	2	0	65	4	2
	对外投资（次）	4	1	0	34	14	1
	融资历史（次）	0	0	1	8	0	0
	失信人信息（次）	0	2	0	0	0	0
	股权质押（次）	0	2	0	3	0	2
牌照状态		被注销	被注销	被注销	正常	正常	正常

注：数据来源为笔者根据天眼查和启信宝数据整理而得。

从表7可以看到，与模型结论和前文分析一致，被注销企业典型代表SXLH公司、CSSL公司和BJJG公司从注册成立至取得牌照的时间天数都在900～1800天，相对合规企业较短；与此同时，媒体报道数普遍较低，新闻动态数量集中于110～370条。所获商标数几乎都为0个，少数企业有注册商标但也为个位数，远远少于合规企业。

其中，SXLH公司牌照注销原因为支付业务设施不符合相关标准及安全要求，支付业务连续性风险大。前文分析指出第三方支付产业链竞争中，商标传递着重要信号，软件著作权也对企业的长期持续发展有重要作用。从SXLH公司的注册商标数和软件著作数来看，公司对于业务创新及发展并不重视，在实际运行中，公司将核心业务系统外包，且外包服务合同已过期。虽然公司2012年就已获得支付牌照，新闻报道数却一直较少，这也进一步传递了企业并不注重业务的拓展与长期发展的信号。

CSSL公司与BJJG公司也都在成立不久后就取得支付牌照，但从媒体报道数较低和注册商标的个数为0这两个关键指标，可以看出两家企业同样并不注重支付业务本身的运营发展，与此同时，在关键指标都远远低于其他企业的情况下，公司的工商变更次数却相对较高，从实际运行情况来看，两家公司的确存在着未经

⑥ 案例牌照信息来自中国人民银行官网公布的各批非银行支付机构《支付业务许可证》续展决定，企业违规事实及注销原因来自金融时报2017年7月2日的报道，详见http://www.financialnews.com.cn/jg/dt/2017 07/t20170702_120213.html。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

核准擅自变更控股股东、擅自变更主要出资人、违规变更股权结构等重大事项，这些都属于变相转让《支付业务许可证》的行为，严重违反了央行的相关规定。

相比之下，对于合规企业，获牌时间变量值都比较高，普遍在2000天以上。媒体报道数则分两个极端，要么非常高，要么很低（低于110条），这些都与前文分析一致。从YLSW公司与GZYL公司来看，两家企业受媒体关注度很高，注册商标数与软件著作数都远远高于其他企业。实际运营情况中，公司注册了金融物管、科学仪器等支付业务相关的商标，开发了诸如网络账户管理软件、发卡管理软件、自助缴费软件等许多软件，以促进保障第三方支付业务的顺利开展与持续运营。JSGX公司则是合规企业中的另一类代表，业务比较集中单一，所以受媒体关注度较低，但同时也意味着企业的负面报道很少。虽然新闻报道少，但是注册商标、软件著作等关键指标依旧体现了企业为了保障支付业务正常运行的投入与付出。这些案例都进一步佐证了本文的研究结论。

（三）预警机制实施及风险监管建议

综上所述，对于央行来说，目前对支付业务许可证的申请条件设置仍有待进一步完善与优化。目前关于支付牌照的申请考核条件中只对企业的注册资本和获牌时间做出明确规定⑦，而忽略了最重要的变量——媒体报道数和企业注册商标数，因而并不能对企业资质进行真正有效的筛选。因此，本文提出基于随机森林的第三方支付违规风险预警机制实施框架，如图4所示。

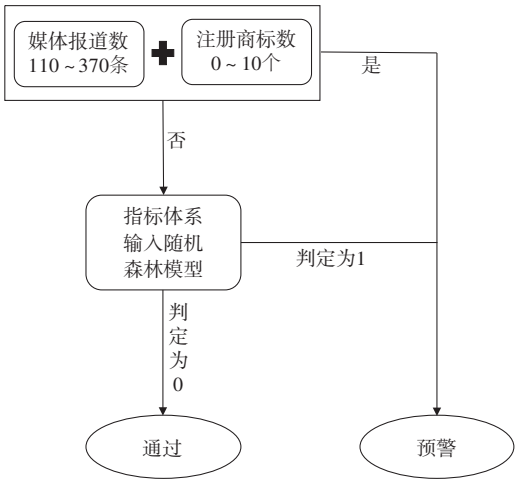


图4　风险预警机制实施框架图

基于所构建的风险预警机制，本文提出如下监管建议：

（1）参考本文所设计的预警指标体系，完善支付业务许可证的申请考核条件。除了企业注册时间，还应该将企业的媒体报道数、注册商标数、软件著作数和经营异常情况、工商变更次数、投资方来源、对外投资等变量都纳入申请考核条件中，要求企业提交相关资料数据证明。

（2）在评判相关企业资质时，首先应重点关注企业的媒体报道数和注册商标数这两个变量，对于媒体报道数在110～370之间的企业，注册商标数非常低甚至接近于0的企业，应设置预警，谨慎给予支付牌照或者直接拒绝。在此基础上，再进一步关注注册时间、软件著作数、经营异常等信息。

（3）关于企业注册资本的规定可以适度放宽，虽然央行在《支付业务许可证》申请要求里有相关规定：企业应有符合规定的注册资本最低限额，但是实际该变量对企业违规经营的影响并不大。央行应重点关注建议（2）所提到的几个变量，在此基础上结合企业注册资本等变量来共同分析。

（4）对于媒体报道数和注册商标数达标的企业，将风险预警指标体系中所有相关指标量化后，运用随机森林模型对企业是否会违规做提前判定，根据模型预测结果，进一步考虑斟酌。对于被判定违规的企业尤其应该审慎考量，从而真正有效地实施对于企业的“健全的组织机构、内部控制制度和风险管理措施”的申请要求。

三、结论

合理有效地监管第三方支付，对支付产业生态的发展至关重要。尤其是随着市场准入限制的进一步放开，数家机构排队申请第三方支付牌照的背景下，建立完善的违规风险预警机制，完备牌照发放和监管机制，能够从根源上进行风险预警，促进支付产业的健康发展。

本文首先基于经典风险理论并结合第三方支付行业的特殊性，构建风险预警指标体系，进而结合机器学习中的随机森林算法，提出了基于随机森林模型的风险预警机制。随后，以获得支付牌照的271家企业为样本，通过对比传统Logistic回归模型与随机森林模型，发现Logistic回归模型在评估第三方支付企业违规行为

⑦ 中国人民银行关于支付业务许可证的申请条件要求详见中国人民银行官网：http://www.pbc.gov.cn//tiaofasi/ 144941/144957/2845832/index.html。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

时有明显的局限性。而随机森林模型判定的准确率大大提高,且一类错误和二类错误比率都大幅降低(二类错误率为0)。这验证了本文所提出的第三方支付违规风险预警机制的有效性。

最后,通过分析每个特征变量在随机森林模型中的贡献程度,得到评价指标重要性的度量值,评估各个属性变量在分类中所起的作用。本文发现媒体报道和注册商标数是最重要的两个变量,而这两点在央行目前对申请牌照企业的考核条件中都没有体现。关于目前申请审核条件中的企业的注册资本要求,本文研究发现其重要程度相对较低,央行日后可适度放宽该项规定。

为了真正有效地对企业资质进行筛选,本文建议央行进一步完善目前对于支付牌照的发放考核条件,在发放牌照时,将企业的媒体报道数、注册商标数、注册时间、软件著作数和经营异常情况等都纳入申请考核条件中,要求企业提交相关资料证明,首先观察企业的媒体报道数和注册商标数是否满足要求,进而通过随机森林模型来评估相关企业的违规风险作为参考,从而降低整个支付行业的风险。

参考文献

[1] 杨彪. 中国第三方支付有效监管研究[M]. 厦门:厦门大学出版社, 2013.

[2] 胡娟. 第三方支付技术与监管[M]. 北京:北京邮电大学出版社, 2016.

[3] YANG Y, LIU Y, LI H, et al. Understanding perceived risks in mobile payment acceptance[J]. Industrial Management & Data Systems, 2015, 115(2): 253-269.

[4] 修永春. "网联"时代第三方支付的三元监管模式探析[J]. 上海金融, 2018(11): 87-91.

[5] XU D, TANG S, GUTTMAN D. China's campaign-style Internet finance governance: Causes, effects, and lessons learned for new information-based approaches to governance[J]. Computer Law & Security Review, 2019, 35(1): 3-14.

[6] 严凌. 我国第三方支付有效监管的研究[J]. 财经界: 学术版, 2018(14): 6-7.

[7] 李晶. 我国第三方支付的风险与防控研究[D]. 杭州: 浙江大学, 2018.

[8] AGRAWAL A, CHADHA S. Corporate governance and accounting scandals[J]. Journal of Law & Economics, 2005, 48(2): 371-406.

[9] ERICKSON M, HANLON M, MAYDEW E L. Is there a link between executive equity incentives and accounting fraud?[J]. Journal of Accounting Research, 2010, 44(1): 113-143.

[10] 冯旭南, 陈工孟. 什么样的上市公司更容易出现信息披露违规——来自中国的证据和启示[J]. 财贸经济, 2011(8): 51-58.

[11] 杨军. 关于国有商业银行信用风险成因与识别的研究[D]. 北京: 清华大学, 2003.

[12] 权小锋, 肖斌卿, 尹洪英. 投资者关系管理能够抑制企业违规风险吗? ——基于A股上市公司投资者关系管理的综合调查[J]. 财经研究, 2016(5): 15-27.

[13] 王晓春. 小微企业风险测度、不良容忍度及其风险控制研究[J]. 金融发展研究, 2012(11): 7-11.

[14] 张淼. Logistic模型构建期货行业风险预警系统研究[J]. 经济研究导刊, 2018(20): 140-142.

[15] 涂艳, 王翔宇. 基于机器学习的P2P网络借贷违约风险预警研究——来自"拍拍贷"的借贷交易证据[J]. 统计与信息论坛, 2018, 33(6): 69-76.

[16] 林成德, 彭国兰. 随机森林在企业信用评估指标体系确定中的应用[J]. 厦门大学学报(自然科学版), 2007(2): 199-203.

[17] 彭国兰, 林成德. 基于随机森林的企业信用评估模型[J]. 福州大学学报(自然科学版), 2008(S1): 153-156.

[18] 于焕杰, 杜子芳. 基于随机森林的企业监管方法研究[J]. 管理世界, 2017(9): 180-181.

[19] 赵晓媚. 上市公司经营风险甄别研究——基于大数据机器学习[J]. 金融管理研究, 2018(2): 153-176.

[20] 向尚, 邹凯, 蒋知义, 等. 基于随机森林的智慧城市信息安全风险预测[C]. 第十八届中国管理科学学术年会论文集. 西安: 中国优选法统筹法与经济数学研究会, 2016: 266-270.

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

[22] ZOU L, CAO K D, WANG Y. Media coverage and the cross-section of stock returns: The Chinese evidence[J]. International Review of Finance, 2019, 19(4): 707-729.

[23] 徐珊, 黄健柏. 媒体治理与企业社会责任[J]. 管理学报, 2015, 12(7): 1072-1081.

[24] 李秀玉, 李倩雯, 史亚雅. 媒体监督对社会责任报告鉴证的影响研究[J]. 经济与管理评论, 2019, 35(6): 81-92.

[25] 贾兴平, 刘益, 廖勇海. 利益相关者压力、企业社会责任与企业价值[J]. 管理学报, 2016, 13(2): 267-274.

[26] 李培功, 沈艺峰. 媒体的公司治理作用: 中国的经验证据[J]. 经济研究, 2010(4): 14-27.

[27] 罗进辉. 媒体报道的公司治理作用——双重代理成本视角[J]. 金融研究, 2012(10): 153-166.

[28] 孔东民, 刘莎莎, 应千伟. 公司行为中的媒体角色: 激浊扬清还是推波助澜?[J]. 管理世界, 2013(7): 145-162.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[29] 周开国，应千伟，钟畅．媒体监督能够起到外部治理的作用吗？——来自中国上市公司违规的证据[J]．金融研究，2016(6)：193-206.
[30] 游家兴，吴静．沉默的螺旋：媒体情绪与资产误定价[J]．经济研究，2012(7)：141-152.
[31] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.
[32] CEFIS E, MARSILI O. A matter of life and death: Innovation and firm survival[J]. Industrial and Corporate Change, 2005, 14(6): 1167-1192.
[33] 肖兴志，何文韬，郭晓丹．能力积累、扩张行为与企业持续生存时间——基于我国战略性新兴产业的企业生存研究[J]．管理世界，2014(2)：77-89.
[34] 吴芃，卢珊，杨楠．财务舞弊视角下媒体关注的公司治理角色研究[J]．中央财经大学学报，2019(3)：51-69.
[35] 应千伟，闵昊婧，邓可斌．媒体关注的市场压力效应及其传导机制[J]．管理科学学报，2017，20(4)：32-49.
[36] 冯晓青．我国企业知识产权运营战略及其实施研究[J]．河北法学，2014，32(10)：10-21.
[37] 潘镇，鲁明泓．中小企业绩效的决定因素——一项对426家企业的实证研究[J]．南开管理评论，2005，8(3)：54-59.
[38] 李维安，李建标．股权、董事会治理与中国上市公司的企业信用[J]．管理世界，2003(9)：103-109.

Research on the Third-party Payment Violation Risk Early-warning Based on Random Forest

Fang Ruonan, Luo Pinliang
(School of Management, Fudan University, Shanghai 200433, China)

Abstract: For the various violation phenomena existing in the third-party payment industry, a violation risk early-warning mechanism based on Random Forest are proposed. A set of index system for violation risk early-warning is firstly built, then the risk early-warning mechanism based on random forest algorithm which comes from machine learning methods is proposed. A sample of 271 companies with paid licenses is used to verify the effectiveness of the proposed early warning mechanism for violation risks. By comparing the discrimination rate of Logistic model and Random Forest model, the research finds that both type I error and type II error of Random Forest model are much lower, and Random Forest model has high accuracy of up to 99. 01% when predicting. Finally, through the analysis of important variables, specific application measures and corresponding risk supervision suggestions are put forward.
Keywords: third-party payment license; risk early-warning; random forest model; risk regulation

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net