

PPP 社会资本参与度预测研究： 基于 14 038 个项目的大数据分析

张经阳^{1,2}, 沈俊鑫¹, 李晓颖¹, 沈冰亮¹

(1. 昆明理工大学管理与经济学院, 云南 昆明 650093;

2. 北部湾大学经济管理学院, 广西 钦州 535000)

摘要:提升 PPP 社会资本参与度,对激发有效投资具有重要意义。考虑到影响社会资本参与 PPP 项目因素的多重复杂性,以 CPPPC 库 14 038 个 PPP 项目为研究对象,选取项目本身、地方政府、市场环境、宏观经济四个维度 61 个特征变量,构建基于集成学习的 PPP 项目社会资本参与度预测模型,并探究不同特征组合对社会资本参与度预测性能的影响。实验结果表明:经蜻蜓算法完成特征降维有效提升了模型预测准确性,ADASYN + XGBoost 模型的预测性能最佳;从特征组合来看,项目本身、地方政府以及宏观经济三种特征组合下的预测性能优于其他特征组合,其中项目本身对社会资本参与 PPP 的影响最为显著,其次是地方政府,市场环境的加入反而会降低除项目本身外其他特征组合下的模型准确性。地方政府应重点从提高财政透明性、提升清廉程度、优化公私合作机制等方面吸引社会资本积极投资。

关键词:政府与社会资本合作;社会资本参与度;集成学习;XGBoost;特征组合

中图分类号:F840.612;F282

文献标识码:A

文章编号:1005-0566(2023)06-0079-13

Research on PPP social capital participation prediction: Based on big data analysis of 14038 projects

ZHANG Jingyang^{1,2}, SHEN Junxin¹, LI Xiaoying¹, SHEN Bingliang¹

(1. Faculty of Management and Economics, Kunming University of Science and Technology, Kunming 650093, China;

2. Faculty of Economics and Management, BeiBu Gulf University, Qinzhou 535000, China)

Abstract: Enhancing the participation of PPP social capital is of great significance to stimulate effective investment. Select 61 characteristic variables in four dimensions of the project itself, local government, market environment, and macroeconomics, build a prediction model of social capital PPP participation based on an ensemble learning algorithm, and explore the influence mechanism of different feature combinations on social capital participation in PPP. The research results show that: 1) Dragonfly algorithm effectively improving the accuracy of the prediction models with feature dimensionality reduction. 2) The prediction model based on ADASYN + XGBoost achieves the best performance under the combination of the characteristics of the project itself, local government and macroeconomics. 3) Among them, the project itself has the most significant impact on social capital participation in PPP, followed by local government; the

收稿日期:2023-01-05 修回日期:2023-06-06

基金项目:国家自然科学基金项目(71964018);云南省省院省校合作项目(SYSX201911);云南产业发展研究项目(2022Z06)。

作者简介:张经阳(1969—),男,贵州遵义人,博士,北部湾大学经济管理学院高级经济师,硕士生导师,研究方向为大数据应用。通信作者:沈俊鑫。

addition of market environment will reduce the accuracy of the model under other characteristic combinations except the project itself; local governments should focus on attracting active investment from social capital by improving fiscal transparency, improving integrity, and optimizing public-private cooperation mechanisms.

Key words: public-private partnership; social capital participation; ensemble learning; XGBoost; feature crosses

随着城镇化的发展,社会公众对公共基础设施服务需求日益渐增,但承担公共基础设施建设融资功能的地方政府融资平台存在过度举债、融资渠道单一、运营效率偏低等问题^[1]。在这一背景下,国家出台多项政策推广、规范公私合作模式(public-private partnership, PPP),从而推进公共基础设施投资主体多元化,解决基础设施建设面临的资金短缺问题,推进我国 PPP 高质量可持续发展^[2]。但由于 PPP 本身具有的参与主体成分复杂、投资规模大、投资回收期长、盈利空间小以及投资风险大等特点^[3-5],使得社会资本尤其是民间资本参与 PPP 项目的积极性较弱,不少项目由于再融资难导致停工或提前终止,“落地率低”、有效投资不足、地方政府财政压力骤增、相关诉讼事件频发等问题也逐渐暴露^[2]。虽然从总体看来,政府更有能力为私营部门参与公共基础设施发展创造有利环境,但 PPP 项目的成功实施极大地依赖于社会资本的参与^[6-7]。社会资本股权占比反映了其参与 PPP 项目的积极性^[1]。据财政部政府和社会资本合作中心(China Public-Private Partnership Center, CPPPC)数据显示:在 14 038 个 PPP 项目中,落地项目社会资本股权比例

分布范围为 50% ~ 100%,且股权比例超过 80% 的占多数,但存在行业、区域异质性(见图 1)。

通过构建基于集成学习的 PPP 社会资本参与度预测模型,达到对社会资本参与情况进行预测预警的目的,一方面可以将关注重点转移到社会资本参与度较低的项目上,提醒相关政府部门及时采取激励措施来提升社会资本的参与度,保证项目的顺利实施。另一方面可以根据可能出现的参与结果对 PPP 项目进行风险评估,把 PPP 项目建设过程中因社会资本参与不足导致的项目停工、终止等问题造成的公私双方的时间和成本损失最小化。为了实现这一目标,本文比较了 3 种不同的集成学习模型来预测 PPP 项目中社会资本参与度的可能性。首先使用蜻蜓算法实现特征降维,然后应用 ADASYN 算法对社会资本参与度样本数据进行非平衡样本集处理,最后利用交叉验证和网格搜索的方式对比评估了 RF(random forest)、GBDT (gradient boosting decision tree)、XGBoost(extreme gradient boosting)3 种模型的预测性能,并将项目本身、地方政府、市场环境、宏观经济四个维度的特征分成 9 个特征组合,探究不同特征组合下模型的预测效果。

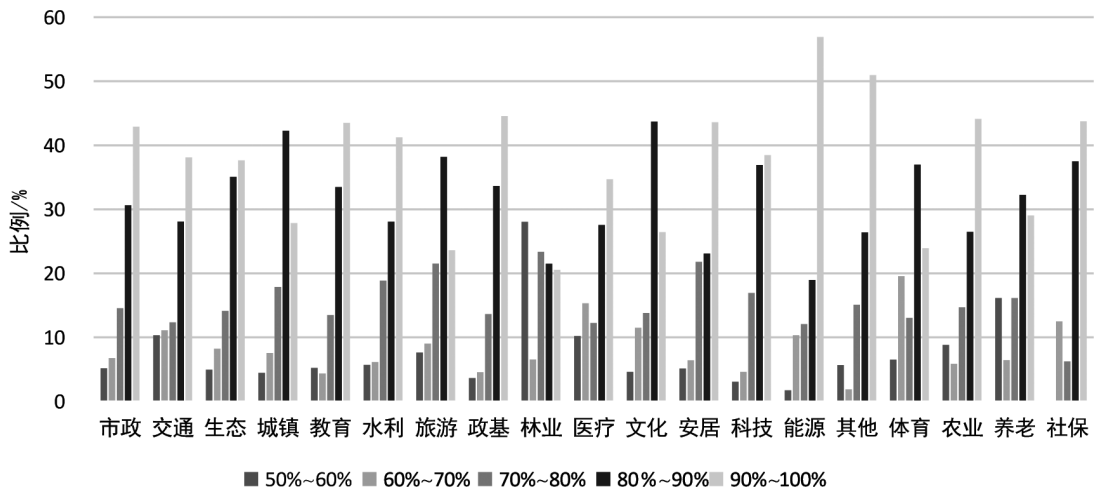


图 1 PPP 各行业社会资本参股比例

一、文献综述及特征选取

(一)文献综述

如何引导社会资本积极参与 PPP 项目也成为国内外学者的研究重点:一方面,项目相关的合作机制是影响社会资本参与 PPP 项目的主要因素,项目期限、项目类别、付费方式、回报机制等均会对社会资本参与 PPP 项目产生影响^[8]。另外,社会资本作为逐利主体,参与 PPP 的主要目标是实现利润最大化。因此,会着重考虑项目带来的收益。具体而言,PPP 项目的盈利能力^[9]包括投资回报率^[10]、合理利润率^[11]以及激励机制^[12],对于吸引社会资本参与 PPP 项目至关重要;合理的风险分配方式则能够实现项目管理过程中的风险最小化,实现利益共享,从而吸引社会资本投资^[11,13]。其次,政府作为 PPP 项目的发起人,其选择偏好^[11,14]、扶持^[1,14]以及担保^[15]政策、腐败程度^[12,16-17]、政府级别^[18]、相关 PPP 经验^[9,12]、财政实力^[19-20]、财政透明度^[21]、管理能力与公信力^[22-23]等均会影响社会资本参与 PPP 项目的积极性。再次,宏观经济稳定性^[12]以及经济发展水平^[18,24]也是影响私人资本参与 PPP 项目的主要因素。一般来讲,经济发展欠佳的地区对私人投资者的吸引力往往较低^[19],但是随着宏观经济状况的改善,政府能够吸引更好的供应商和更多的融资^[9]。最后,市场的存在则使得 PPP 有了更大的盈利的空间而吸引社会资本的投资^[22]。就目前来看,学者多倾向于探究市场规模^[9,12]、市场需求^[24]以及市场化程度^[1]对社会资本参与 PPP 项目的促进作用。

由于现有学者在探索影响社会资本参与 PPP 项目因素中主要使用计量经济学模型,其受限于数据维度的影响,且随着维度增加,会导致非参数估计的准确性下降,从而陷入维度灾难^[25];传统的计量经济学模型主要用于因果推断无法实现有效预测,现有的社会资本参与度研究仍停留在影响因素分析阶段。大数据时代,机器学习作为一种可靠的预测分析方法,广泛应用于管理、金融、医疗等领域。目前,机器学习在 PPP 领域中的应用

仍处探索阶段:Owolabi 等^[26]、Wan 等^[27]分别使用随机森林和支持向量机方法对 PPP 风险进行预测;Wang 等^[28]使用机器学习模型预测 PPP 项目的成败并确定导致 PPP 合同失败的因素;沈俊鑫等^[2]基于集成 LightGBM-Blending 算法,对中国 PPP 项目可融资性进行评价,证明了机器学习方法可有效应用于不同行业 PPP 可融资性评估。随着机器学习技术的不断深入,传统的单一模型已经不能满足现有的预测需求,集成学习模型通过综合多个基学习器可以提升模型的泛化能力并达到最佳的预测性能,越来越受研究者青睐。

(二)特征选取

PPP 作为合作伙伴关系重要治理工具,涉及公共利益、国家责任以及管理、融资、效率、治理、技术等。提升社会资本参与 PPP 意愿,应基于 PPP 项目治理、公共治理及双边匹配理论。社会福利最大化视角下社会资本参与意愿受项目属性、付费模式、运作方式、项目结构等内部特征,以及宏观环境、市场化水平、公共财政以及政府治理能力等外部特征影响。

本文将影响社会资本参与 PPP 项目的主要因素归纳为地方政府、宏观经济、市场环境、项目本身 4 个维度,其中前 3 个维度属于外部因素,具体特征如表 1 所示。

除表 1 中显示的特征指标以外,本文从 CPPPC 项目库中爬取了 PPP 项目相关信息,基于项目数据的完整性最终选取的项目本身因素如表 2 所示。表 1 和表 2 中的“*”均表示构造特征,其中类别型变量的构造方法包括:行业壁垒参照冯净冰等^[1]的方法,根据 PPP 项目所处行业划分为低壁垒以及高壁垒行业,如市政工程属于高壁垒行业,旅游、养老属于低壁垒行业;政府级别根据 PPP 项目发起的政府单位,划分为省级、市级、县级 3 个等级,涵盖 31 省(自治区、直辖市)、346 个地市、2 286 个县区,宏观经济、人口规模、地方政府为项目所在县区数据,其他市场环境为项目所在地市数据;地区经济水平参照《中国区域经济高质量发展研究报告》,根据 PPP 项目所在省份划分为要素缺乏、中低质量、中等质量、中高质量、高质量 5 个经

表 1 PPP 项目外部因素

所属	特征名称	特征解释
宏观经济	GDP	地区生产总值 ^[9,12,28]
	GDPGrowthVolatility	近 5 年 GDP 增速的标准差
	PerCapitaGDP	人均 GDP ^[1,8-9,18-20,24]
	PPI	生产者价格指数
	CPI	消费者价格指数 ^[21]
	RPI	零售物价指数
	PPIGrowthVolatility	生产者价格指数近 5 年浮动的标准差
	CPIGrowthVolatility	消费者价格指数近 5 年浮动的标准差
	RPIGrowthVolatility	零售物价指数近 5 年浮动的标准差
	PerCapitaIncome	居民人均可支配收入
	PerCapitaExpenditure	居民人均消费支出
	ecnoDivision *	地区经济水平
市场环境	Population	人口规模;潜在消费需求 ^[9,24,29]
	MarketizationProcess	市场化进程相对指数 ^[1,17,29]
	GovernmentMarketRelationship	政府与市场关系指数 ^[1,17,29]
	NonStateOwnedEconomy	非国有经济发展指数 ^[1,17,29]
	ProductMarket	产品市场指数 ^[1,17,29]
	FactorMarket	要素市场指数 ^[1,17,29]
地方政府	GeneralPublicBudgetRevenue	一般公共预算收入;财政实力 ^[1,19]
	GeneralPublicBudgetExpenditure	一般公共预算支出 ^[1]
	FinancialTransparency	财政透明性 ^[21]
	Corruption	腐败程度 ^[12,16-17]
	GovConsumExpenditure	政府消费支出 ^[24]
	GovSupport	政府对基础设施建设的支持力度 ^[12,24]
	LocalGovernmentDebtBalance	地方政府债务余额 ^[9]
	FiscalSelfFinancing	财政自给率 ^[20]
	DeficitRatio	赤字率 ^[12]
	LiabilityRate	负债率 ^[9]
	govLevel *	政府级别 ^[19]
	projExperience *	政府 PPP 项目经验 ^[9,12]

济级别。数值型变量的构造方法包括:审批耗时表示从项目发起到本级政府财政承受能力审核通过所隔天数;社会资本招募耗时表示项目本级政府批准拟签署的 PPP 合同与项目发起时间所隔天数;政府 PPP 项目经验表示项目发起单位之前负责的 PPP 项目个数;所耗时间与所处阶段比例表示项目从发起至今耗费天数与项目目前所处阶段的比值;合同透明性由 CPPPC 项目库中项目必填信息完整程度以及非必填信息完整程度按照 8:2 的比例计算得出。

二、数据来源及预处理

(一)数据来源及特征分析

CPPPC 项目管理库收录了我国所有通过物有所值评价和财政承受能力论证的 PPP 项目,并按要求公开了相关项目信息,为研究 PPP 项目提供了数据源。本文采用爬虫技术获取 CPPPC 项目库中的项目相关信息,基于数据完整性原则选取了 2013—2022 年 14 038 个项目作为研究对象,宏观经济、市场环境、地方政府相关特征数据来源于万德数据库、统计年鉴、政府工作报告以及《中国分省份市场化指数报告》《中国市级政府财政透明度研究报告》等。研究发现,项目数据的特征表现如下。

1. 标签离散化

本文依据冯净冰等^[1]的方法,采用注册资本中社会资本的占比,即股权占比衡量 PPP 项目中

表 2 PPP 项目本身因素

特征名称	特征解释	特征类型	特征名称	特征解释	特征类型
investCount	项目总投资\项目规模	数值型	industryRequiredName	所属行业	类别型
capitalProject	资本金	数值型	cooperationTerm	合作期限	数值型
socailInvestAmount	社会资本总投资	数值型	operateMode	运作方式	类别型
userPayAmt	使用者付费收入	数值型	returnMode	回报机制	类别型
capitalRecepts	资本性收益	数值型	socialPurchaseWay	采购方式	类别型
discountInputPay	政府配套投入现值	数值型	tradeBarriers *	行业壁垒	类别型
discountRiskPay	政府自留风险现值	数值型	geogDivision *	所属地理区域	类别型
discountOperationSubsidy	政府运营补贴现值	数值型	timeInterval *	项目耗时与所处阶段比例	数值型
beforeTaxIRR	全投资税前内部收益率	数值型	evalRate	物有所值指数	数值型
afterTaxIRR	全投资税后内部收益率	数值型	socialBearRiskCount	社会资本承担风险种类	数值型
discentRate	折现率	数值型	socIRR	社会资本收益率	数值型
transferRiskCost	社会资本承担风险现值	数值型	cnaValue	竞争性中立调整现值	数值型
vfmValue	物有所值现值	数值型	approvalTime *	审批耗时	数值型
netProfit	净利润现值	数值型	capitalRecruitmentTime *	社会资本招募耗时	数值型
constructionInterest	建设期利息现值	数值型	projStatus	所处阶段	类别型
projRiskCost	项目风险现值	数值型	contrTransparency *	合同透明性	数值型

社会资本的参与情况,其中财政部财金[2014]156号《关于规范政府和社会资本合作合同管理工作的通知》的PPP合同指南中明确规定,政府的持股比例应当低于50%,故社会资本的实际股权范围为50%~100%。社会资本在PPP项目中股权占比呈离散化分布,本文根据先验经验将标签值以股权80%为界限划分为两个区间,其中社会资本股权占比高于80%的项目标记为社会资本具有高参与度,社会资本股权占比低于80%的项目标记为非高参与度。

2. 数据噪声大

在国家政策的支持下,大量项目分批进入CPPPC管理项目库中,由于项目开始时间不同,导致项目信息完整性参差不齐,存在关键特征数据缺失等情况;人为向PPP管理项目库中输入数据,存在一定的输入错误,造成了数据异常的情况;不同特征数据的尺度相差大,在分类过程中会影响模型收敛速度以及准确性。为了消除噪声信息,需要对数据进行相关处理。

(二)数据预处理

1. 异常值检测

异常数据的存在不利于观察实验对象的变化规律,并对结果分析造成一定干扰,对异常值进行检测与处理可以保证原始数据的可靠性,提高模型准确性。基于明树数据关于PPP项目投资回报率的统计结果,本文首先剔除全投资税前内部收益率(afterTaxIRR)、全投资税后内部收益率(beforeTaxIRR)、社会资本收益率(socIRR)、折现

率(discountRate)4列中小于等于0或大于等于15的值,然后利用箱型图检测保留数据的异常状态,最终的检测与处理结果如图2所示;对于物有所值指数(evalRate)则保留0~100范围内的值。另外,根据国发[2019]26号《国务院关于加强固定资产投资项目资本金管理的通知》所示PPP项目最低资本金比例为20%,且项目资本金经验值维持在50%以内,故资本金(capitalProject)范围设置在项目总投资的20%~50%;其余与PPP项目投入以及花费金额相关特征属性需保证值大于等于0,并且最大值不能超过项目总投资50%。

2. 缺失值填充

不同特征列的缺失值需要采用不同的方法,相关数据缺失情况如图3所示。具体来讲,before Tax IRR、after Tax IRR与soc IRR通过所属行业以及项目规模等字段填充均值;transfer Risk Cost、discount Risk Pay先按照项目类型分类,采用不同项目类型中相关风险占项目投资比例的众数计算并填补缺失值;construction Interest根据项目运营方式的不同,将不存在建设期项目填充为0,其他缺失值利用项目运营方式、项目类型、项目规模等字段填充均值;discount Operation Subsidy根据项目回报方式不同,将回报方式为使用者付费的项目填充为0,其他缺失值根据不同的回报方式按照项目规模填充均值;discount Rate根据项目类型所属区域以及发起时间填充均值;cna Value填补方法与风险类字段相同;capitalReceipits列利用均值填充。

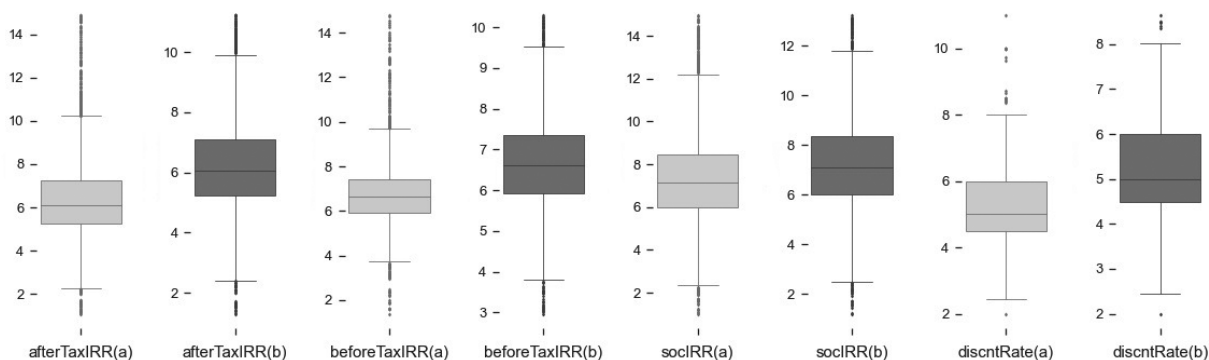


图2 箱图异常值检测与处理

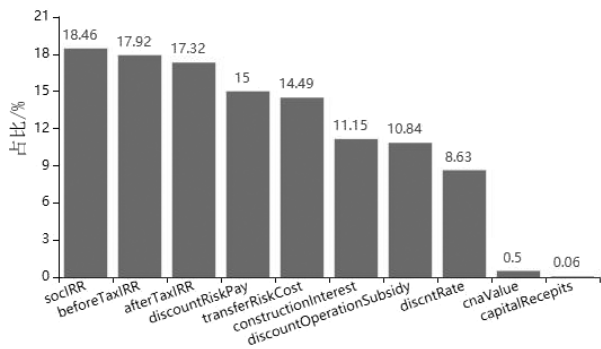


图3 数据缺失占比

3. 类别型变量

具体的类别型变量的数据描述如表3所示,不同类别型变量编码依据如下:项目所处阶段、行业壁垒、政府级别、地区经济水平均为有序变量,故依据变量值的顺序使用序列填充的形式编码。财政部相关文件将PPP运作方式分为委托运营(OM)、管理合同(MC)、转让—运营—移交(TOT)、改建—运营—移交(ROT)、建设—运营—移交(BOT)和建设—拥有一运营(BOO)等6种。本文考虑到PPP模式应用的复杂性将其重新归纳为三大类,即外包类、特许经营类、私有化类。相关研究发现随着私有化程度的增高,地方政府让渡给社会资本的权利就越大,项目对于社会资本的吸引力也逐渐增加^[30]。因此,项目运作方式根据私有化程度使用序列填充的形式编码,私有化程度越高,编码越大。袁竞峰等^[31]分析了不同回报机制下社会风险爆发情况,认为使用者付费、可行性缺口补助、政府付费三种机制下社会风险等级会依此降低。对于社会资本来讲,风险等级越低越容易获取稳健的投资回报,从而激发其投资积极性。因此,回报机制根据社会风险等级进行编码,社会风险等级越高,编码越小。财政部发布的财库〔2014〕215号《政策和社会资本合作项目政府采购管理办法》中明确规定PPP项目采购方式包括公开招标、邀请招标、竞争性谈判、竞争性磋商和单一来源采购。其中公开招标的对投标人的要求相对宽松;邀请招标由于采购项目的特殊性,虽然中标率高但投资效益不佳;竞争性谈判和竞争性磋商对于投标人的具体要求和适用范围相似;单一来源采购是为应对紧急情况而采用的采购

方式,由于涉及金额巨大一般不采用此种方式。因此,采购方式按照社会资本进入PPP项目的难度:单一来源采购>邀请招标>竞争性谈判=竞争性磋商采购方式>公开招标的顺序进行编码,其中,公开招标占比高于2/3,且5种采购方式的公开透明性存在逐渐上升趋势,透明性低的采购方式更容易发生违法违规操作^[32],因此编码就越小。所属地理区域以及所属行业的取值不存在先后优先级关系,因此使用one-hot编码处理。

表3 类别变量数据描述

变量	变量值	频数	百分比/%	编码
projStatus (项目所处阶段)	准备阶段	245	5.29	1
	采购阶段	1 042	22.49	2
	执行阶段	3 346	72.22	3
tradeBarriers (行业壁垒)	低壁垒	760	16.4	1
	高壁垒	3 873	83.6	2
govLevel (政府级别)	县级	3 574	77.14	1
	市级	977	21.09	2
	省级	82	1.77	3
ecnoDivision (地区经济水平)	要素缺乏区	969	20.92	0
	中低质量区	396	8.55	1
	中等质量区	1 669	36.02	2
	中高质量区	1 054	22.75	3
	高质量区	545	11.76	4
operateMode (运作方式)	其他	414	8.94	0
	外包类	4	0.09	1
	特许经营类	4 172	90.05	2
	私有化类	43	0.93	3
returnMode (回报机制)	使用者付费	171	3.69	1
	可行性缺口补助	3 008	64.93	2
	政府付费	1 454	31.38	3
socialPurchaseWay (采购方式)	单一来源采购	40	0.86	1
	邀请招标	12	0.26	2
	竞争性磋商、竞争性谈判	687	14.83	3
	公开招标	3 894	84.05	4

4. 数据标准化

由于数据集中各特征的单位尺度相差较大,在分类过程中,高数值水平的特征可能会削弱低数值水平特征的作用,从而降低模型精度。因此,为了保证模型预测结果的可靠性,需要消除各特征之间因数量级不同造成的影响。数据的标准化处理通过将数据按比例缩放当一定的区间,可以把原始数据转化为无量纲、无数量级差异的标准化数值,达到去除数据单位限制的目的。本研究采用标准分数(Z-score)来标准化原始数据,标准化后的数据符合标准正态分布,Z-score的计算公式如下:

$$x = \frac{x - \mu}{\sigma} \quad (1)$$

式中, μ 为所有样本数据的均值; σ 为所有样本数据的标准差。

三、模型构建与结果分析

(一) 模型构建流程

PPP 社会资本参与度大数据预测模型构建流程如图 4 所示, 共分为 4 个模块: 数据采集、数据处理、模型构建与模型评估与结果分析。

1. 数据采集分为项目外部因素数据以及项目本身因素两个部分, 将项目外部因素数据根据 PPP 项目发起年份的滞前一期以及所处地区与项目本

身因素匹配整合成完整的数据集合。

2. 数据处理包括标签、异常值、缺失值、类别变量以及数据标准化处理, 同时还根据文献以及相关经验构造了少量特征列, 例如审批耗时、合同透明性、政府的 PPP 经验等。

3. 模型构建中首先使用蜻蜓算法实现特征选择, 然后使用 ADASYN 算法处理非平衡样本集问题以提升模型性能, 最后利用交叉验证和网格搜索的方式选择最优参数下的最佳模型。

4. 模型评估与结果分析通过对比相同特征不同模型以及相同模型不同特征的预测结果, 选出最佳性能模型下的最优特征组合。

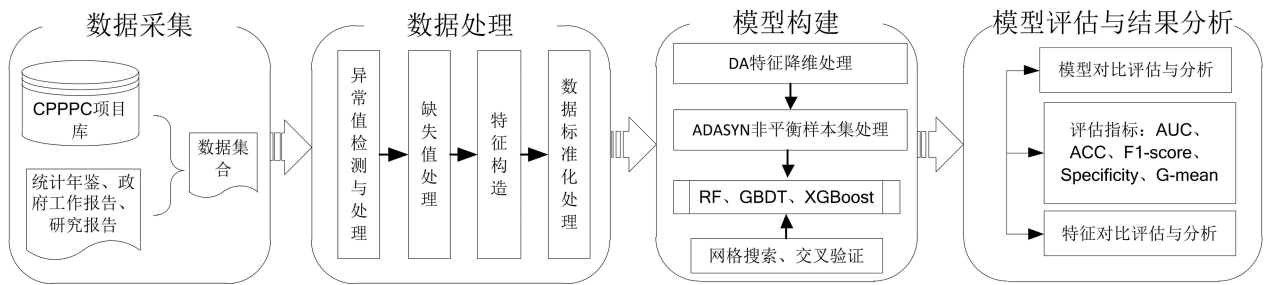


图4 预测模型

(二) DA 特征降维

蜻蜓算法 (dragonfly algorithm, DA) 是由 Mirjalili 在 2016 年提出, 具有结构简单、易于实现且鲁棒性强的特点, 主要用于解决小样本、高维度、不平衡的数据预测问题^[33]。此算法借鉴蜻蜓群体静态群体和动态群体两种行为模式, 在静态群体中, 蜻蜓会分为几个不同的子群体在不同区域中搜寻食物, 对应着群智能算法中的全局搜索; 动态群体中, 群体则会聚集成一个大群体沿着一个方向迁徙, 对应着群智能算法中的局部开发。影响社会资本参与 PPP 意愿包括 61 个特征, 需要进行特征降维, 具体步骤如下。

1. 对于给定特征集 $F = \{f_1, f_2, \dots, f_N\}$, N 为特征数量, 则特征子集为 $S = \{s_1, s_2, \dots, s_n\}$, 其中 $s_i = \{0, 1\}$, $i = 1, 2, \dots, n$, $s_i = 0$ 表示第 i 个特征 f_i 不被选择, 反之则被选择。

2. 创建初始化蜻蜓种群数量 X_i , 其中 $i = 1, 2, \dots, n$, ΔX 为步长向量, X 为位置向量, 初始化个体单位步长 ΔX_i 。

$$\Delta X_{t+1} = (sS_i + aA_i + cC_i + fF_i + eE_i) + \omega \Delta X_t \quad (2)$$

式中, w, s, a, c, f 分别为五种行为权重, ω 为惯性权重; S_i, A_i, C_i, F_i, E_i 分别蜻蜓的避撞、结队、聚集、觅食和避敌行为向量, t 为当前的迭代次数, X_t 为当前 t 代种群的位置。

3. 计算转换函数并更新蜻蜓个体位置。当有邻近蜻蜓时 $X_{t+1} = X_t + \Delta X_{t+1}$; 当周围无邻近蜻蜓时, 蜻蜓采用随机游走策略: $X_{t+1} = X_t + Levy(d) \times X_t$ 。

式中, ΔX_{t+1} 为 $t+1$ 代位置更新步长, X_{t+1} 为 $t+1$ 代种群的位置; d 为位置向量维数, $Levy$ 为设定的飞行函数。

(三) ADASYN 非平衡数据处理

采样优化作为解决非平衡数据集问题有效方法之一, 通过不同的抽样方式调整数据集样本分布以实现样本集平衡, 主要包括过采样或欠采样。ADASYN 是一种不平衡学习的自适应综合采样方法^[34], 通过引入密度分布函数 Γ_i , 计算每个少数

类样本需要生成的新样本数量,然后再利用 SMOTE 算法合成新样本。具体而言,社会资本非高参与度类样本数据的合成步骤如下。

1. 计算样本数据的不平衡度 d :

$$d = S_{min}/S_{maj} \quad (3)$$

式中, $d \in (0,1]$, S_{min} 为社会资本非高参与度类样本总量, S_{maj} 为社会资本高参与度类样本总量。

2. 计算需要合成的社会资本非高参与度类样本数量 G :

$$G = (S_{maj} - S_{min}) \times \beta \quad (4)$$

式中, $\beta \in [0,1]$; 当 $\beta = 1$ 时,合成数据后的社会资本非高参与度样本和高参与度样本数量达到平衡。

3. 计算每个社会资本非高参与度类样本 x_i 需要生成的样本数量 g_i :

$$g_i = \Gamma_i \times G, \quad \Gamma_i = \frac{\Delta_i/K}{Z} \quad (5)$$

式中 Δ_i 为 x_i 样本的 K 个近邻中多数类样本的数量, Z 为规范化因子。

4. 利用 SMOTE 算法合成新的社会资本非高参与度类样本 x_{ij} : 在每个待合成的社会资本非高参与度类样本 x_i 的 K 个近邻中随机选取若干样本,对于每一个选中的样 x_j , 在 x_i 、 x_j 之间的连线上随机选一点作为新合成的社会资本非高参与度类样本 x_{ij} , 样本 x_{ij} 的合成公式为:

$$x_{ij} = x_i + rand(0,1) \times |x_i - x_j| \quad (6)$$

(四) 集成学习模型

集成学习模型通过将多个单一机器学习算法集成融合的方式,可以有效地降低单一模型的偏差和方差,从而使模型在准确度、泛化能力、鲁棒性等方面得到优化。其中, Bagging 和 Boosting 是两种常见的集成学习方法。在 Bagging 方法中,使用随机抽样的方法获取到多个相互独立的训练集,每个训练集并行的训练模型,各基学习器之间互不干扰且具有相同权重,可以有效降低方差,代表算法随机森林。在 Boosting 方法中,各基学习器采用同一个训练集,通过串行训练的方式使得新生成的基学习器都用来修正上一轮训练中的误

差,各基学习器通过其分类误差来分配权重。这种方法通过迭代的方式可以降低模型偏差,代表算法 GBDT。因此为了更好地对比两种集成方法,本文分别选取了其代表算法进行研究,另外 XGBoost 在 GBDT 的基础上引入了正则化和特征子采样,在降低模型偏差的基础上可以防止模型过拟合,达到降低方差的目的。

1. RF

RF 是一种 Bagging 集成学习算法,在分类问题中通过投票的方式统计训练的 K 棵决策树的预测结果作为最终的结果。通过引入两个“随机性”:“随机选择样本集”和“随机抽取特征子集”,以防止模型陷入过拟合状态,提升模型抗噪能力。单棵决策树的生长机制为:①对于给定的社会资本参与度训练集样本 N ,利用自助法重采样技术从训练集中抽取 N 个样本作为每颗决策树的训练样本;②对于每个社会资本参与度样本拥有的 M 个属性,在决策树的节点需要分裂时,从 M 个属性里随机选取 m 个属性,且 $m \leq M$,然后从这 m 个属性里使用信息增益策略选择最优属性进行分裂;③每棵树都按照步骤②进行分裂,直到该节点到达叶子节点时分裂结束,整个分裂过程无需剪枝。重复上述决策树的成长过程直至建立起 K 棵决策树,可以得到如下的分类决策:

$$H(x) = \arg \max_y \sum_{k=1}^K I[h_k(x) = y] \quad (6)$$

式中, h_k 表示第 k 棵决策树, $I[h_k(x) = y]$ 为示性函数,当 $h_k(x) = y$ 为真时 $I(\cdot) = 1$, 为假时 $I(\cdot) = 0$ 。

2. GBDT

GBDT 作为一种以决策树为基分类器的 Boosting 模型,通过多轮迭代降低残差来提高最终分类器的精度,要建立社会资本参与度的 GBDT 预测模型,首先要初始化学习器,对于第一颗分类决策树 $F_0(x)$ 。其初始化规则为:

$$F_0(x) = \log \frac{P(y = 1 | X)}{1 - P(y = 1 | X)} \quad (7)$$

式中, $P(y = 1 | X)$ 表示对于给定的训练集 X 中,社会资本高参与度所占的比例。

然后进行迭代优化,即通过使用梯度提升方

法分析残差来拟合模型,在第 k 次迭代中(共迭代 K 次),对于每一个样本,其损失值的负梯度(残差估计值)计算公式为:

$$r_{k,i} = -\frac{\partial L(y_i, F(x_i))}{\partial F(x)} = y_i - F(x_i) \quad (8)$$

在确定了残差估计值之后,利用 CART 回归树拟合数据,则第 k 棵树对应叶子区域 $R_{k,j}$ 的最佳拟合值计算公式为:

$$c_{k,j} = \frac{\sum_{x_i \in R_{k,j}} r_{k,i}}{\sum_{x_i \in R_{k,j}} (y_i - r_{k,i})(1 - y_i + r_{k,i})} \quad (9)$$

式中, $j = 1, 2, \dots, J_k$, J_k 表示第 k 棵决策树的叶子节点个数。经过迭代优化后的学习器可以表示为:

$$F_k(x) = F_{k-1}(x) + \sum_{j=1}^{J_k} c_{k,j} \cdot \eta \quad (10)$$

式中, η 为学习率。最终可以得到基于 GBDT 的社会资本参与度预测模型:

$$F_K(x) = F_0(x) + \sum_{k=1}^K \sum_{j=1}^{J_k} c_{k,j} \cdot \eta \quad (11)$$

3. XGBoost

XGBoost 是一种 Boosting 集成学习算法,基本思想是通过不断地进行特征分裂来生成树,利用每轮中学习的新树,去拟合上一轮模型预测值与实际值之间的残差,通过迭代上述过程对目标函数进行优化。在给定的社会资本参与度数据集中,对于第 t 棵树, $f_t(x_i)$ 为第 t 棵树的预测结果,则第 i 个社会资本参与度样本 x_i 的预测参与度为 $\hat{y}_i^{(t)}$:

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (12)$$

社会资本参与度模型训练的目标函数为:

$$Obj^{(t)} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k),$$

$$where \Omega(f) = YT + \frac{1}{2} \lambda \|\omega\|^2 \quad (13)$$

正则项 $\Omega(f_k)$ 中 Y, λ 表示惩罚系数, T 表示给定一颗树的叶节点数, $\|\omega\|^2$ 表示每棵树叶子节点上的输出分数的平方, $\Omega(f_k)$ 主要是用来控制决策树的复杂度,有利于防止过拟合,从而提高模型的泛化能力。

(五) 结果与讨论

本文分为两个阶段测量和验证预测模型的评估效果。第一阶段,对比多种模型的预测结果,选出性能最佳的模型;第二阶段,验证多种特征组合下的模型最佳预测效果,获取最优的特征组合。

1. 模型对比评估

对于不同的模型均使用网格搜索(GridSearchCV)实现自动化调参以确定最优参数(见表4),并采用以下5项评价指标评估模型性能:准确率(ACC)、特异度(Specificity)、F1-score、G-mean、和 AUC 值,性能对比结果如表5所示。

表4 算法参数设置

RF		GBDT		XGBoost	
参数名	参数值	参数名	参数值	参数名	参数值
n_estimators	107	learning_rate	0.19	learning_rate	0.1
max_depth	7	n_estimators	50	n_estimators	420
max_features	12	max_depth	14	max_depth	5
min_samples_leaf	6	min_samples_split	15	min_child_weight	2
min_samples_split	60	subsample	0.9	subsample	1

表5 各分类模型性能对比

算法名称	ACC	Specificity	F1-score	G-mean
RF	0.866 3	0.1	0.927 2	0.316 2
GBDT	0.856 4	0.3	0.918 8	0.534 8
XGBoost	0.856 4	0.266 7	0.919 2	0.505 8
ADASYN + RF	0.865 1	0.840 9	0.864 7	0.865 5
ADASYN + GBDT	0.909 6	0.916 1	0.907 7	0.909 7
ADASYN + XGBoost	0.920 8	0.920 5	0.918 4	0.920 8

图5显示了分别用多种模型预测 PPP 项目中社会资本的参与度。ROC 曲线下面积即 AUC 值,介于 0—1 之间,越接近 1 表示模型预测的真实性越高。从图中可以看出,使用非平衡数据算法处理过的数据训练出来的模型性能更好,且 ROC 曲线下面积都大于 0.85,说明非平衡数据算法 + 集

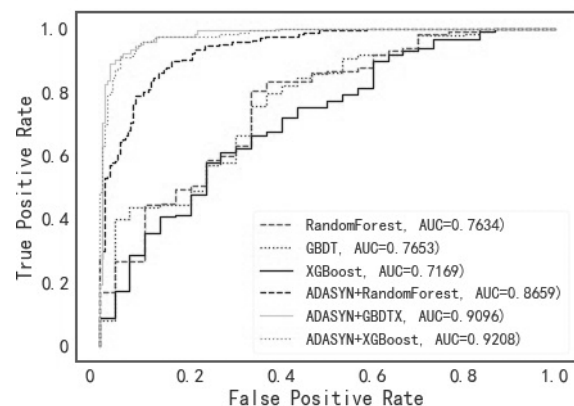


图5 ROC 曲线对比

成学习模型可以有效的完成 PPP 项目社会资本参与度的预测,结合表 5 可以看出 ADASYM + XGBoost 模型的综合评价最优,因此利用 ADASYM + XGBoost 模型进行第二阶段的预测研究。

2. 特征对比评估

通过上述模型对比评估并选出拥有最佳性能的模型 ADASYM + XGBoost 后,对不同特征组合的预测结果再次进行对比分析,选出表现最优的特征组合。由于本文是从 4 个角度选取的特征:项目本身(P)、宏观经济(E)、市场环境(M)、地方政府(G),因此选取以下 9 种特征组合方式进行 PPP 项

目的社会资本参与度预测研究:①项目本身(P);②项目本身和地方政府组合(PG);③项目本身和市场环境组合(PM);④项目本身和宏观经济组合(PE);⑤项目本身、地方政府以及市场环境组合(PGM);⑥项目本身、地方政府以及宏观经济组合(PGE);⑦项目本身、市场环境以及宏观经济组合(PME);⑧地方政府、市场环境以及宏观经济组合(GME);⑨全部特征组合(PGME)。预测结果仍使用上述五个指标进行评估,如图 6 所示,展示了不同评估指标与 1 之间的误差值,因此阴影面积越小说明预测效果越好。

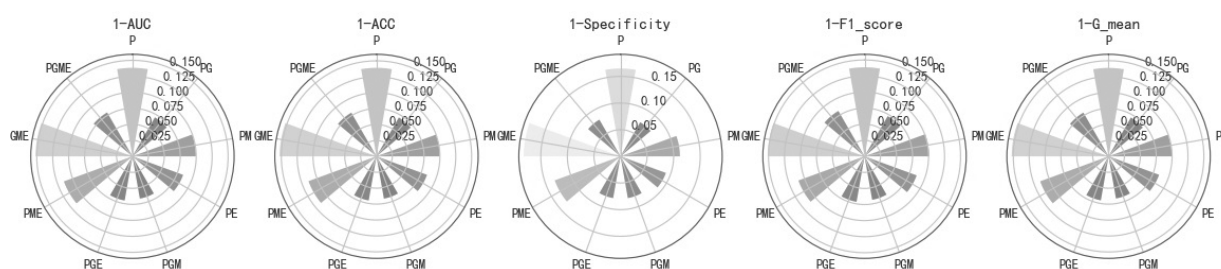


图 6 不同特征组合的预测结果对比

从图 6 中可以得出如下结论:①由组合 P 和 GME 对比其他特征组合可知,只考虑项目本身因素,或者不考虑项目因素时,模型性能都较其他特征组合低,因此在研究影响社会资本参与 PPP 项目问题时不能单纯地只考虑项目本身因素或者项目外部因素,因为社会资本在投资 PPP 项目时会综合考虑项目本身属性以及包含地方政府、宏观经济、市场环境等因素在内的地区属性^[1,24]。②对比组合 P、PG、PM 以及 PE 可知,在考虑了项目本身因素的情况下,政府因素的加入较市场因素以及经济因素更能提高模型性能,说明政府因素会显著影响社会资本参与 PPP 项目的积极性。另外,经济因素以及市场因素也在一定程度上对社会资本参与 PPP 项目产生影响,而且经济因素的影响更大;其中政府、经济以及市场相关因素对社会资本参与 PPP 项目的积极因素也在之前学者的研究中得以证实。③由组合 PG 对比组合 PGM、PGE 可知,在项目与政府组合的基础上多加入其他任何一种因素,均能提高模型性能,说明从多个角度考虑 PPP 项目对社会资本的吸引力是有必要的。④对比组合 PE 与

PME 或者 PGE 与 PGME 可知,市场因素的加入会降低模型的预测结果的准确性,可能是以下几种情况导致了这种现象发生:一是市场因素对社会资本参与度预测的影响被政府以及经济因素所覆盖,因为项目与市场因素结合下的模型预测性能优于只有项目本身因素的模型,而且上述结论②也表明市场因素对于社会资本参与 PPP 项目的影响确实较政府因素与经济因素弱。二是数据粒度原因。市场因素的粒度相对于其他因素的粒度过大,政府因素以及经济因素包含的指标多是匹配的地级市指标,基于数据的可得性只收集到了以省为单位的市场相关数据,由于数据粒度大小不同,可能会导致模型性能有所下降。

四、结论与建议

(一) 结论

当前,我国 PPP 模式已经进入高质量发展阶段,社会资本的积极参与能够有效促进 PPP 的可持续发展。PPP 社会资本参与意愿受项目内外部环境众多因素影响,机器学习擅于处理海量、高维数据,突破了传统计量模型在预测技术上的局限

性,具有更加精准的样本外预测能力;集成学习作为机器学习的一个分支,通过组合若干基学习器来达到提升模型性能的目的,集成模型可以有效地防止模型欠拟合以及过拟合,提高模型泛化能力;与此同时,蜻蜓算法通过降维方式提升了集成学习算法的预测性能。通过爬取 PPP 项目内外部多源异构信息,在巨量、高维数据支撑下,集成学习模型可有效预测 PPP 项目中社会资本的参与度,一方面为政府促进社会资本积极参与 PPP 项目提供决策参考,另一方面为社会资本如何选择适合的 PPP 项目提供思路。

本文通过对比不同集成学习模型的预测性能,在最优模型下考察不同特征组合对社会资本参与度的预测影响,研究结果发现:首先,ADASYN 与 Random Forest、BGDT、XGBoost 结合的方式提高了原有模型的性能,其中 ADASYN + XGBoost 的方式在本文使用的数据集上达到了最佳的预测性能,准确率及 AUC 均为 92.08%,说明集成学习算法能够有效地预测 PPP 项目的社会资本参与度。其次,通过对比不同特征组合下的模型准确性,发现项目本身、地方政府、宏观经济是造成社会资本参与 PPP 项目的主要因素,其中项目和政府因素对社会资本参与 PPP 项目的影响最为显著。最后,在只考虑项目本身的情况,市场环境的加入会提升模型准确性;但当地方政府和宏观经济相关指标共同存在时,加入市场环境指标则会降低模型准确性。初步判断上述现象可能是市场环境并非强相关因素或者数据粒度不统一引起的,但是后者还需进一步证实。

本文可能存在以下 3 个方面的边际贡献:第一,丰富了人工智能应用领域,验证了大数据驱动决策范式在 PPP 投资决策的可行性,挑战了模型驱动、理论驱动在 PPP 管理理论发展中的核心地位。近年来,随着数字经济快速发展,CPPPC 中心发布了大量的项目信息,可以从项目内外部特征获得足够的样本数据,并借鉴先进的大数据和人工智能技术,为地方政府、社会资本和金融机构提供同时克服模型驱动变量分布假设要求苛刻、模型选择缺乏灵活性问题。第二,多源、多维、异构

是基于机器学习的 PPP 决策面临的首要难题,本文从项目内外部爬取决策特征信息,涉及项目本身、地方政府、市场环境、宏观经济 4 个维度 61 个特征变量,通过引入 DA 算法及 ADASYN 算法,有效解决了大数据驱动 PPP 投资决策高维、非平衡问题,为解决大数据环境下管理决策“维数灾难”问题提供借鉴和启发。第三,本文基于数据驱动,探究不同特征组合下对社会资本参与度预测性能的影响,挖掘不同特征组合与社会资本参与度的内在关系,发现项目本身对社会资本参与 PPP 的影响最为显著,其次是地方政府,而市场环境的加入反而会降低除项目本身外其他特征组合下的模型准确性,印证了 PPP 作为合作伙伴关系重要治理工具,涉及公共利益、国家责任以及管理、融资、效率、治理、技术等。因此,需要基于双边匹配理论,从微观项目治理、宏观公共治理提出政策建议,助推 PPP 高质量发展。

(二) 建议

随着 CPPPC 中项目信息的不断完善以及国家治理数字化的加速推进,以大数据为支撑的 PPP 管理决策成为新的趋势。政府部门可以通过使用集成学习模型预测 PPP 项目是否对社会资本具有吸引力进而评估项目的可行性,社会资本可以用其来判断 PPP 项目是否值得参与。

由于政府因素对社会资本参与 PPP 项目具有显著影响,而与财政实力相关的因素短时间内无法得到较大提升的方面,因此政府可以通过建立透明的监管机制,提高财政透明性、提升清廉程度、优化营商环境等吸引技术和资金实力雄厚的社会资本参与 PPP 建设。但是 PPP 项目中,地方财政实力是政府偿付能力和履约能力的重要保障,是影响社会资本参与政府付费(或补助)项目的重要因素,财政自给率较低且未来政府支出责任较高地区,为防止形成政府隐性债务,应重点通过降低政府持股比例,增加单位 PPP 投资额对应的土地供给,优化公私合作机制,建立良好伙伴关系,提升社会资本参与 PPP 项目积极性。

PPP 作为一种合作模式,公共部门和社会资本之间是双向选择关系,地方政府应着力提升治理

水平,尽量避免地方政府“选择偏好”,激发民营资本投资热情。丰厚的利润回报是社会资本参与 PPP 项目的首要因素,但 PPP 项目存在投资周期长、投资风险巨大等问题,政府可以通过制定合理的风险分配方式来提升社会资本参与 PPP 项目的参与度。

参考文献:

- [1]冯净冰,章韬,陈钊. 政府引导与市场活力:中国 PPP 项目的社会资本吸纳[J]. 经济科学, 2020(5): 19-20,24-26.
- [2]沈俊鑫,吕佳历,程墙,等. 中国 PPP 项目可融资性差吗:基于集成 LightGBM-Blending 算法[J]. 中国软科学, 2022(1): 50-61.
- [3]牛耘诗,褚晓凌,王守清,等. PPP 项目提前终止下社会资本机会主义行为研究[J]. 系统工程理论与实践, 2019, 39(11): 2784-2791.
- [4]张晓雪,刘继才. 金融机构提前介入 PPP 项目的风险管理激励效应研究[J/OL]. 中国管理科学: 1-12 [2022-12-08]. <https://doi.org/10.16381/j.cnki.issn1003-207x.2021.2643>.
- [5]ZHANG S, CHAN A P C, FENG Y, et al. Critical review on PPP research: a search from the Chinese and international journals [J]. International journal of project management, 2016, 34(4): 597-612.
- [6]龚强,张一林,雷丽衡. 政府与社会资本合作(PPP): 不完全合约视角下的公共品负担理论[J]. 经济研究, 2019, 54(4): 133-148.
- [7]ZHANG X Q. Critical success factors for public-private partnerships in infrastructure development [J]. Journal of construction engineering and management, 2005, 131(1): 3-14.
- [8]曾莉,罗双双. 我国 PPP 实践中民营企业参与度及其影响因素研究:基于 731 个县域样本的实证分析[J]. 软科学, 2020, 34(2): 33-38.
- [9]CALLENS C, VERHOEST K, BOON J. Combined effects of procurement and collaboration on innovation in public-private-partnerships: a qualitative comparative analysis of 24 infrastructure projects[J]. Public management review, 2021(10): 1-21.
- [10]龚军蛟,张敏. 民营企业参与城市公用事业 PPP 项目的影响因素研究[J]. 经济理论与经济管理, 2020(5): 100-112.
- [11]王雨辰,胡轶俊. 民营企业 PPP 项目参与度研究:基于公共性的风险分析[J]. 软科学, 2019, 33(6): 89-94.
- [12]ZHU J B, HERTOOGH M, ZHANG J W, et al. Incentive mechanisms in mega project-risk management considering owner and insurance company as principals[J]. Journal of construction engineering and management, 2020, 146(10): 1-12.
- [13]ZHANG X X, LIU J C. Incentive mechanism and value-added in PPP projects considering financial institutions' early intervention [J]. Journal of construction engineering and management, 2022, 148(4): 1-14.
- [14]薛菁. 政策扶持对民营企业 PPP 项目参与度影响效应探析[J]. 兰州学刊, 2020(4): 138-148.
- [15]沈言言,刘小川. 促进私人部门 PPP 投资的政府担保政策研究:基于中低收入国家 PPP 项目的证据[J]. 财政研究, 2019(5): 33-46.
- [16]YE X, SHI S, CHONG H-Y, et al. Empirical analysis of firms' willingness to participate in infrastructure PPP projects [J]. Journal of construction engineering and management, 2018, 144(1): 04017092.
- [17]王欢明,陈佳璐. 地方政府治理体系对 PPP 落地率的影响研究:基于中国省级政府的模糊集定性比较分析[J]. 公共管理与政策评论, 2021, 10(1): 115-126.
- [18]仇娟东,黄海楠,马赫然. 多边金融机构如何影响 PPP 项目中私人部门的投资额:来自“一带一路”沿线经济体 3 858 个项目的经验证据[J]. 中国软科学, 2021(1): 85-97.
- [19]ZHANG Y. From state to market: private participation in china's urban infrastructure sectors, 1992 - 2008 [J]. World development, 2014(64): 473-486.
- [20]沈言言,李振. 地方政府自有财力对私人部门参与 PPP 项目的影响及其作用机制[J]. 财政研究, 2021(1): 116-129.
- [21]刘畅,郭雪萌,王卓君. 中国式财政分权、财政透明度与地方政府选用 PPP 模式[J]. 经济问题探索, 2020(2): 170-182.
- [22]YANG Y, HOU Y, WANG Y. On the development of public-private partnerships in transitional economies: an explanatory framework [J]. Public administration review, 2013, 73(2): 301-310.
- [23]PERCOCO M. Quality of institutions and private participation in transport infrastructure investment: evidence from developing countries [J]. Transportation research part A: policy and practice, 2014(70): 50-58.
- [24]杨丽花,王喆. 私人资本参与 PPP 项目的影响因素

- 分析:基于亚投行背景下的经验分析 [J]. 亚太经济, 2018 (1): 53-61,146.
- [25] 纪园园, 谢婧青, 李世奇, 等. 计量经济学前沿理论与方法:第四届中国计量经济学者论坛(2020)综述 [J]. 经济研究, 2021, 56(4): 201-204.
- [26] OWOLABI H A, BILAL M, OYEDELE L O, et al. Predicting completion risk in PPP projects using big data analytics [J]. IEEE transactions on engineering management, 2020, 67(2): 430-453.
- [27] WAN A, FANG J. Risk prediction of expressway PPP project based on PSO-SVM algorithm [M]. American Society of Civil Engineers Reston, VA, 2020: 55-63.
- [28] WANG Y, SHAO Z, TIONG R L. Data-driven prediction of contract failure of public-private partnership projects [J]. Journal of construction engineering and management, 2021, 147(8): 04021089.
- [29] 王卓君, 郭雪萌, 李红昌. 地区市场化进程会促进地方政府选用 PPP 模式融资吗:基于基础设施领域的实证研究 [J]. 财政研究, 2017 (10): 54-64,91.
- [30] HU Z B, WU G D, WANG H M, et al. Dynamics in network governance of infrastructure public-private partnerships: evidence from four municipalities of China [J/OL]. Public management review:1-25[2023-05-01]. DOI: 10.1080/14719037.2023.2206844.
- [31] 袁竞峰, 尚东浩, 邱作舟, 等. 不同回报机制下 PPP 项目社会风险涌现机理研究 [J]. 系统工程理论与实践, 2020, 40(2): 484-498.
- [32] 翟磊, 袁慧赞. 政府如何影响 PPP 项目社会资本投资水平:基于 3561 个项目的实证研究 [J]. 甘肃行政学院学报, 2021(4): 97-110,127.
- [33] MIRJALILI S. Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems [J]. Neural computing and applications, 2016, 27(4): 1053-1073.
- [34] HE H, BAI Y, GARCIA E A, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning [C]. IEEE, 2008: 1322-1328.

(本文责编:王延芳)