解码数据宇宙：

数据科学与机器学习的现状

返回目录

研究目标

有几个挑战阻碍了组织成功将机器学习（ML）模型整合到其软件开发生命周期中。 弥合不同技能集之间的差距，处理复杂和大型数据集，管理专用硬件，并确保生产中的可用性、可扩展性和安全性共同延迟了价值的实现，并导致组织瓶颈。

由于对机器学习项目的兴趣和复杂性不断增加，组织需要通过适当规模的治理来提高敏捷性、效率和性能，并降低风险。 组织认识到他们需要明确的数据科学和机器学习策略。 作为这些策略的一部分，MLOps可以提供一种结构化和标准化的方法来开发、部署和维护生产中的ML模型，以实现更大的价值。 为了进一步了解这些趋势，TechTarget的企业战略集团（ESG）对北美（美国和加拿大）的366名与数据科学和机器学习技术和流程相关的专业人士进行了调查，包括可能负责制定战略、评估、购买、构建和管理这些技术的责任。

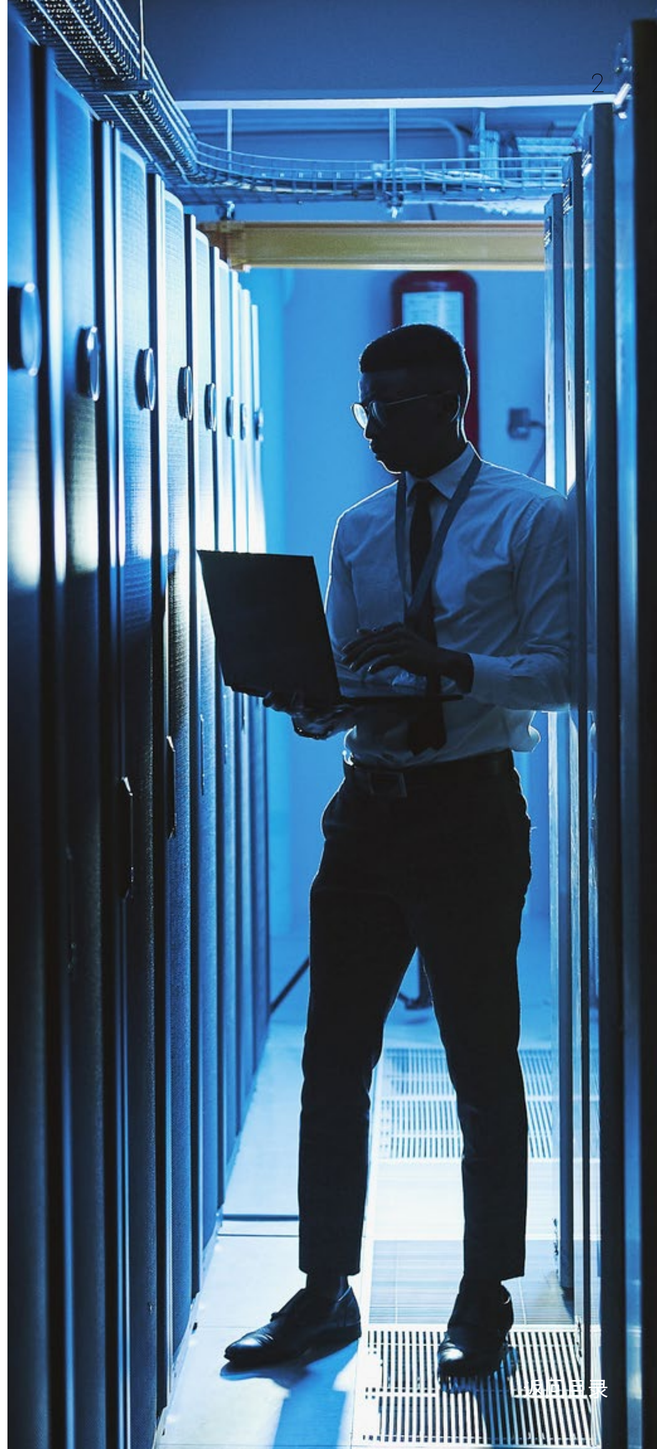
本研究旨在：

识别数据科学和机器学习计划、目标和挑战。

确定组织如何优先考虑解决方案以帮助它们取得成功。

建立通过MLOps实现人工智能的当前状态。

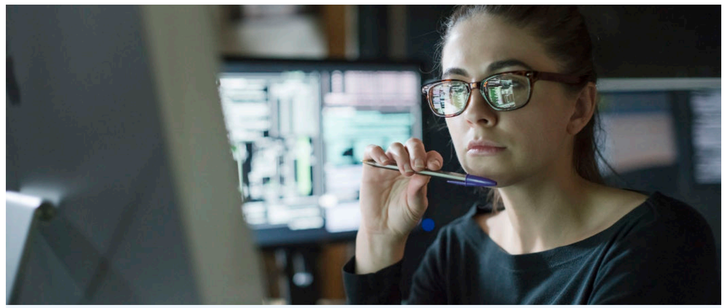
了解不断发展的利益相关者格局，包括团队组成、参与度和增长机会。



返回目录

主要发现

点击关注

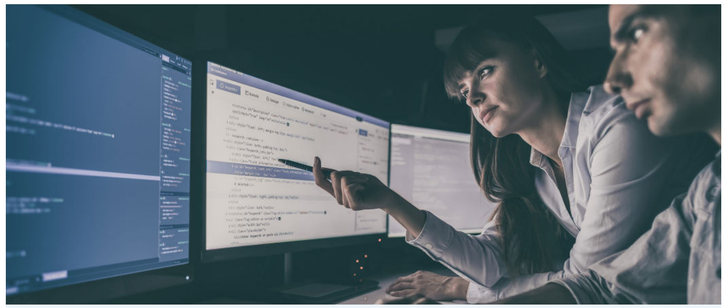


投资指向惊人的增长，但挑战依然存在

第4页

关注点集中在改善数据科学生命周期的早期和后期阶段

第10页



组织提高了将模型转化为生产环境的能力，但仍需要进一步提高效率

第14页

数据科学和机器学习成为一项团队运动，供应商专注于使所有利益相关者能够参与

第17页

投资指向惊人的增长，但挑战依然存在

主要业务目标指向内部

提高运营效率仍然是推动数据科学和机器学习倡议的关键，以实现大多数业务目标。 它不仅赋予组织提高敏捷性、成本效益和客户中心性的能力，还为可持续增长和规模化奠定了基础，在一个日益数据驱动的世界中。 一旦运营达到最佳水平，组织可以更多地关注其他业务要务。 然而，数据科学和机器学习倡议还有望改进产品开发、客户体验、风险管理和其他领域。

数据科学和机器学习倡议的主要业务目标。

66%
提高运营效率

60%
提高产品开发和创新

52%
提升客户体验/提高客户满意度

49%
提高风险管理

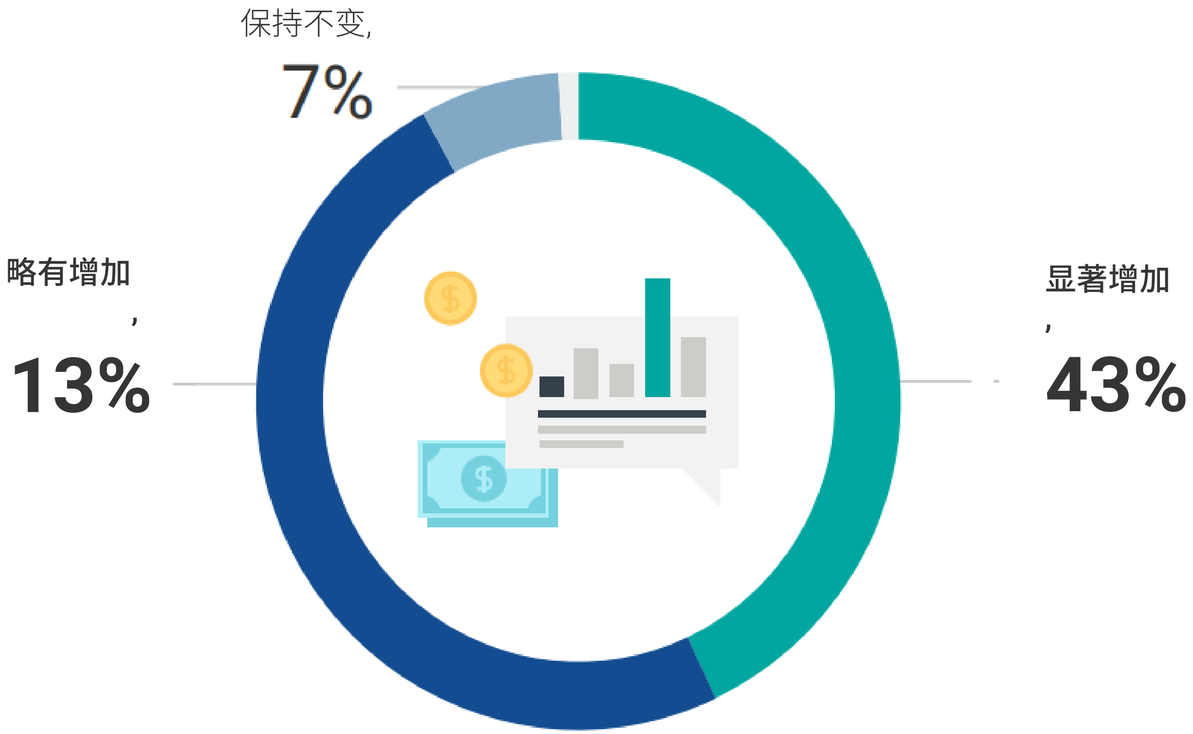
47%
增强决策能力

43%
发现新的商机和/或增加收入

返回目录

“这种增加的投资反映出人们对于**数据科学不仅提高了运营效率，还实现了明智的决策，预测分析和创新产品开发。**”

数据科学和机器学习项目/倡议的预算变化与上一年相比。



预算正在增加

几乎所有（92%）的组织都看到了数据科学和机器学习项目/倡议的预算分配年度增加。 这些预算是相当可观的，近四分之一（24%）的组织计划在未来几年内至少投资100万美元用于与数据科学和机器学习相关的人员、流程或技术。 这种增加的投资反映出人们对数据科学的理解，它不仅增强了运营效率，还实现了明智的决策、预测分析和创新产品开发。 这种财务支持强调了数据科学和机器学习在从庞大而复杂的数据集中提取有价值知识方面所扮演的关键角色，推动组织在数字时代取得成功。

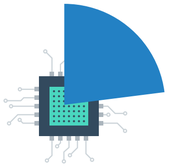
返回目录

在优先考虑数据科学项目时，策略是多样的

愿意牺牲时间来市场并在有限资源下继续前进，突显了组织采取的谨慎乐观的态度。 他们认识到他们不能等待，但也必须确保健壮的模型开发，全面的测试和准确的洞察，以避免潜在的昂贵错误。 这种有意识和计算的方法可以提高长期性能，可靠性和利益相关者的信心，远远超过最初的时间投资。

对数据科学相关项目的优先处理方法

23%
业务影响
（即具有最高潜在业务影响的项目）



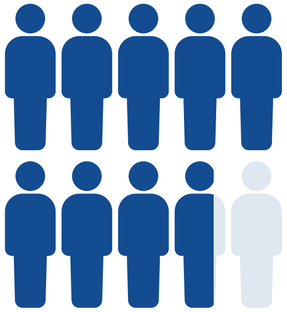
23%
技术复杂性
（即具有最高技术复杂性的项目）

7%
上市时间
（即，具有最短上市时间的项目）

13%
资源可用性
（即，可以使用现有资源完成的项目）

14%
客户反馈
（即，解决客户反馈的项目）

19%
高管领导
（即，优先级由高管领导团队决定的项目）



88%
组织中的大部分人同意
开源对数据科学和机器学习的创新至关重要。

返回目录

衡量数据科学项目影响的艺术

每个数据科学项目都为衡量影响带来了独特的维度。 回应的接近程度证明了数据科学在不同领域中的多样化方法和用例，突显了其变革力量。

由于提高运营效率是数据科学倡议最常见的业务驱动因素，因此它也是最常用于确保这些策略性能的衡量领域。 客户满意度和成本节约也常常被监测以确定这些倡议的影响。

用于衡量数据科学项目/倡议的领域。

53%
提高运营效率

48%
客户满意度

45%
节约成本或创造收入

39%
节约时间

37%
竞争优势

37%
预测准确性

36%
创新潜力

35%
员工满意度/幸福感

26%
社会影响

返回目录

挑战重重

几乎所有（94%）的组织在开发和实施数据科学项目时面临挑战。

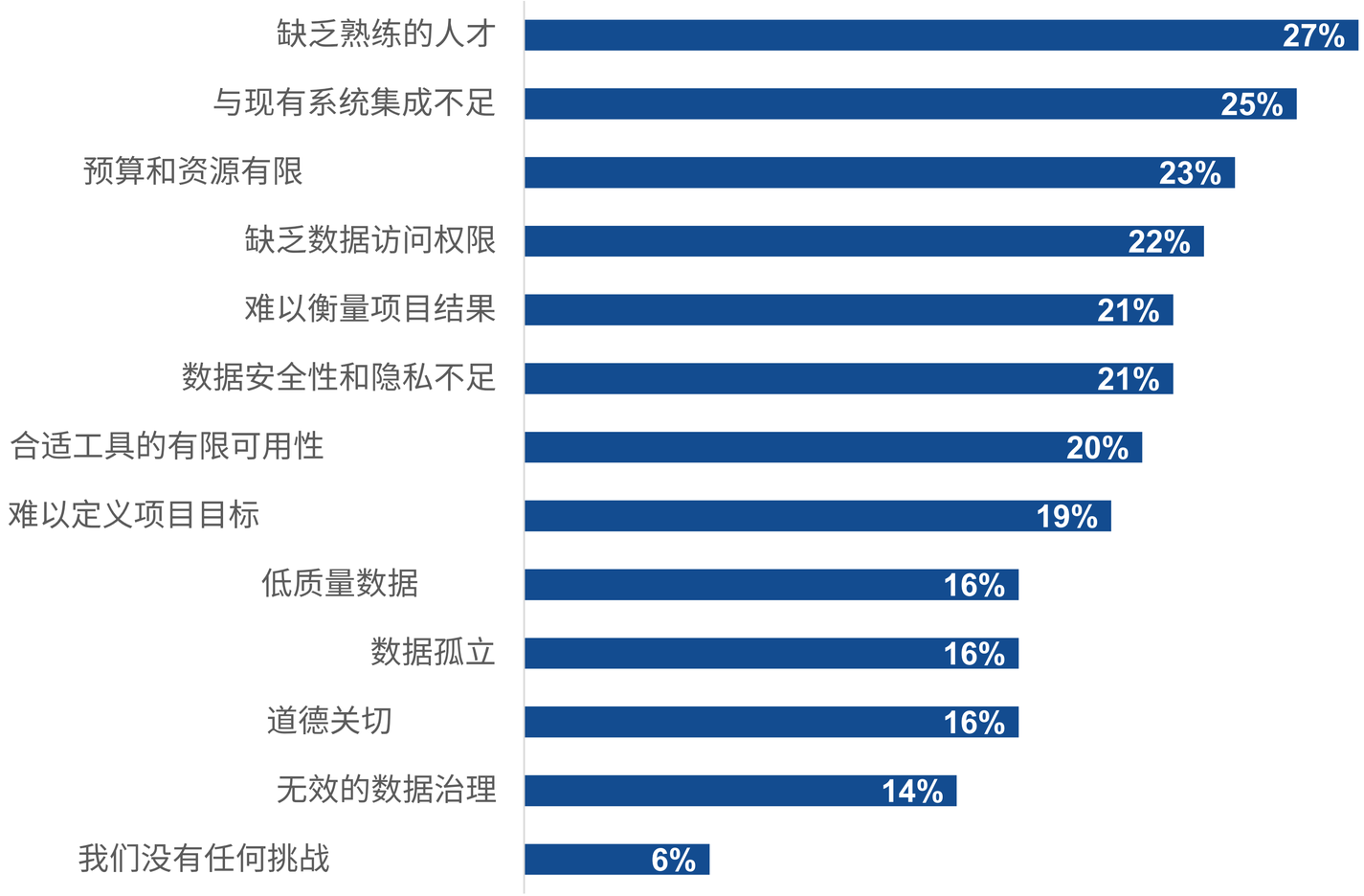
挑战有各种形式和规模：

组织层面：
熟练的人才、预算、确定目标和衡量结果。

数据/环境层面：
与现有系统集成、数据可访问性、有限的工具、低质量数据和数据孤立。

信任层面：
数据安全/隐私、道德关切和数据治理。

在开发和实施数据科学项目中面临的最重要的挑战。

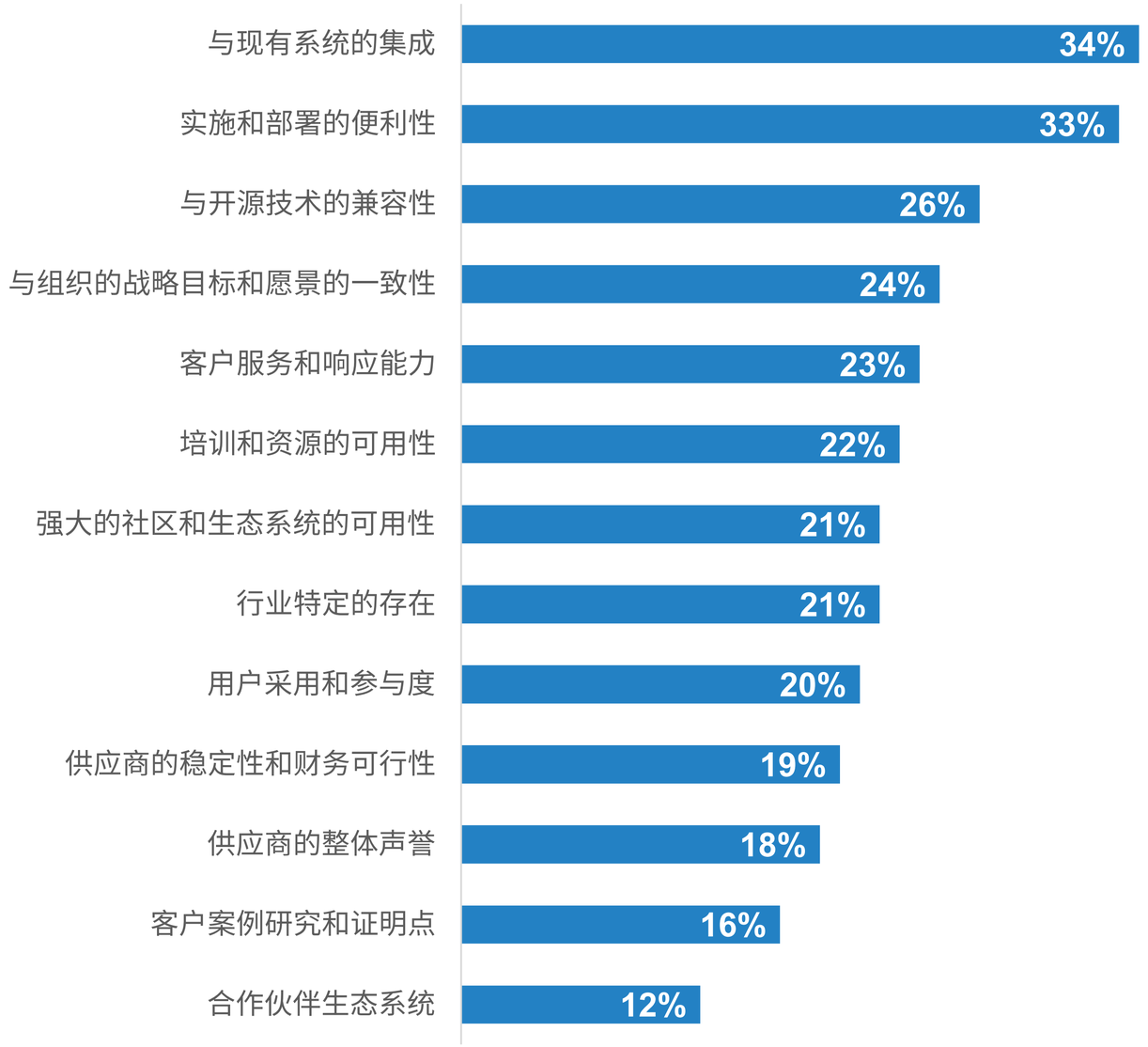


返回目录

关注数据科学生命周期的早期和后期阶段

数据科学与机器学习的现状

在考虑支持数据科学项目时最重要的因素。



在考虑数据科学购买时权衡的因素突出了对集成和简单性的渴望

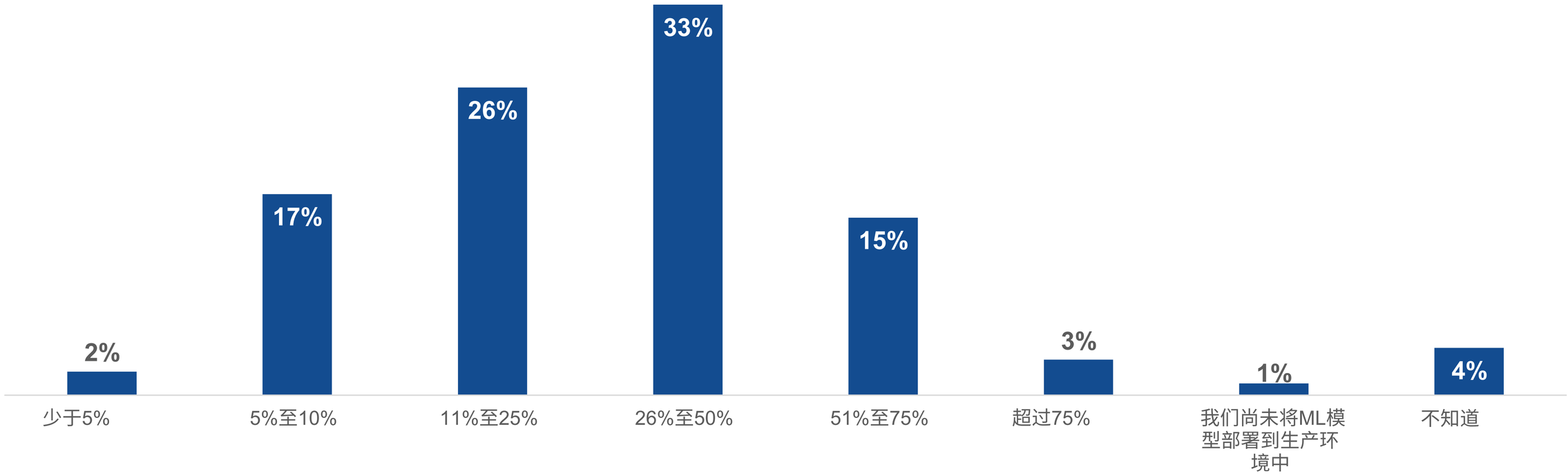
许多组织已经在他们的数据科学和机器学习计划中进行了大规模投资，因此确保他们仍然从这些投资中获得价值至关重要。简化实施和部署突显了组织迅速提升和改善数据生成和数据洞察之间时间的愿望。 还要注意，超过四分之一（26%）的组织考虑与开源技术的兼容性，这可能预示着一个更大的开源部署趋势。

返回目录

将模型迁移到生产环境中仍有很大的改进空间

在过去一年中，组织在改善机器学习模型的操作化和过渡到生产环境方面取得了巨大进展。 通过强大的框架和自动化流程进行模型训练、验证和部署，行业在现有系统中实现了更无缝的集成，同时也实现了更快的迭代。 这种改善成功的根本在于MLOps实践的出现，促进了数据和IT利益相关者之间的合作。 然而，尽管取得了这些改进，组织在将机器学习模型部署到生产环境中的速度仍有很大的提升空间。 例如，45%的组织只有不到25%的模型能够进入生产环境。 在管理模型的整个生命周期中仍然存在挑战，从最初的开发到持续的监控和维护，以应对模型漂移、性能下降、可解释性问题等等。

机器学习模型部署到生产环境中的百分比。

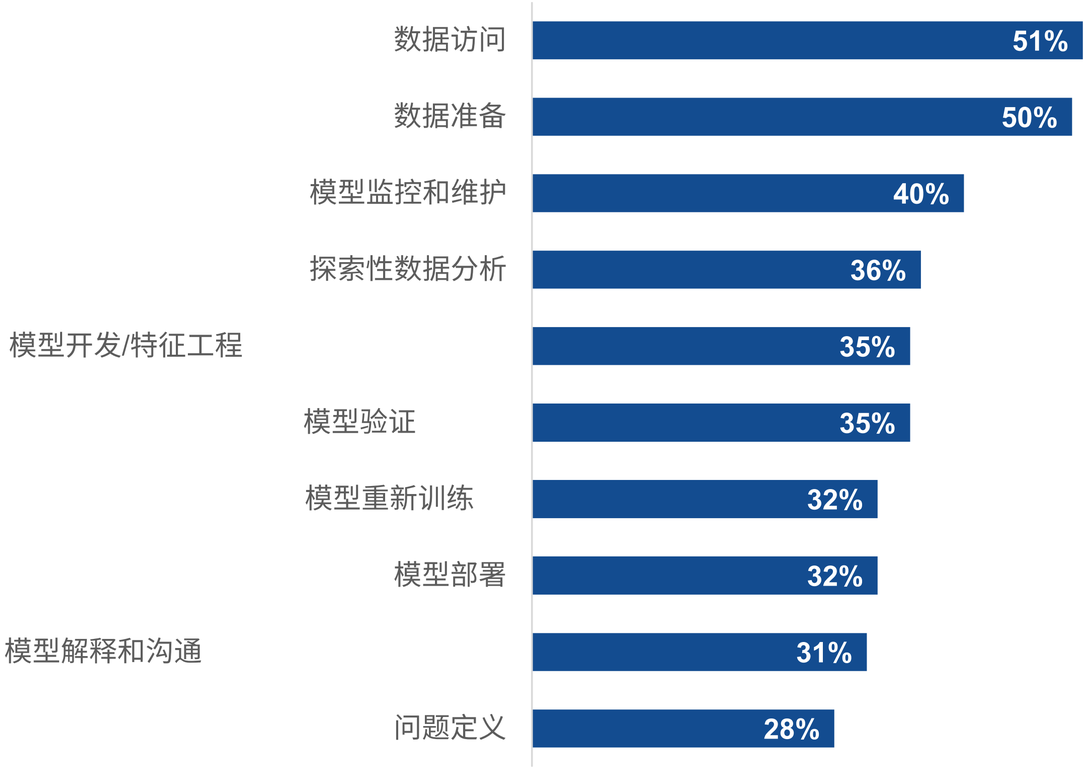


返回目录

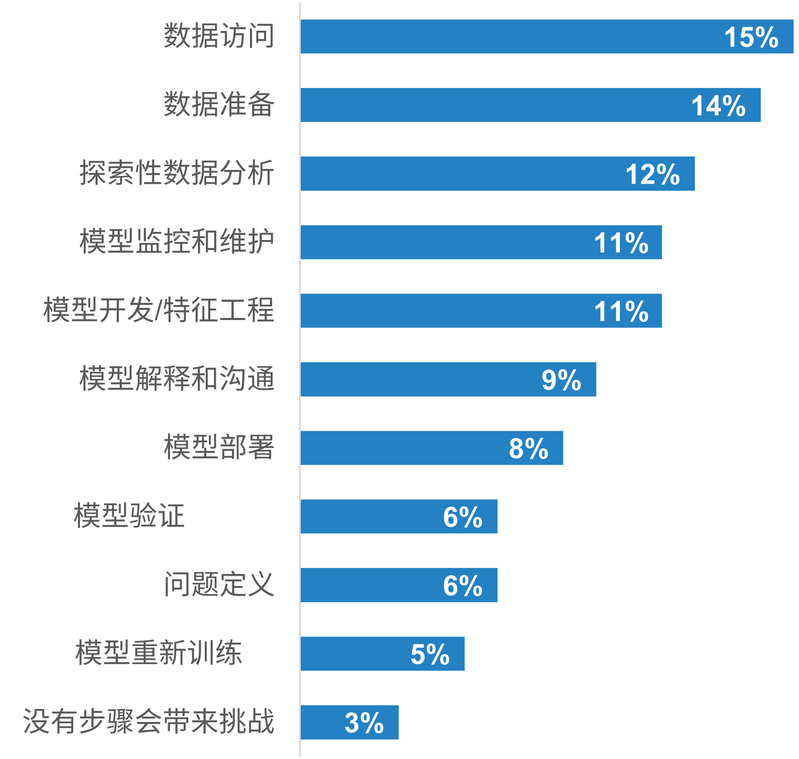
数据的重要性不可低估

数据可访问性和数据准备是相辅相成的。 数据可访问性构成整个数据科学生命周期的基础，不仅强调为什么这是最常见的定期操作，还强调为什么它对组织来说是最大的挑战。 数据准备，包括数据清洗、结构化和转换，是确保后续的分析实验建立在可靠和准确的基础上的必要步骤。

数据科学生命周期中的常规步骤。



最具挑战性的数据科学生命周期步骤。



返回目录

组织提高了将模型转移到生产环境的能力，但需要进一步提高效率

拆解机器学习部署和监控中的挑战

考虑到58%的组织在将模型投入生产过程中有很大的改进空间，即使是最成熟的组织也会遇到挑战。 在将模型集成到现有基础设施中、确保与各种系统的兼容性以及遇到意外的现实世界数据变化时，会出现技术复杂性。 合规和治理挑战会影响可靠性和信任，并引入风险。 操作复杂性会出现，例如随时间保持模型性能和识别/响应故障。 持续监控也带来挑战，例如解决数据漂移和管理模型依赖性，如模型版本控制。

机器学习模型部署和监控的挑战。

35%
管理多个环境的困难

33%
确保符合公司治理政策的困难

33%
检测和响应数据漂移的困难

29%
生产中模型性能不一致

29%
检测和响应模型故障的困难

26%
保留过程效率低下

26%
管理依赖关系的困难

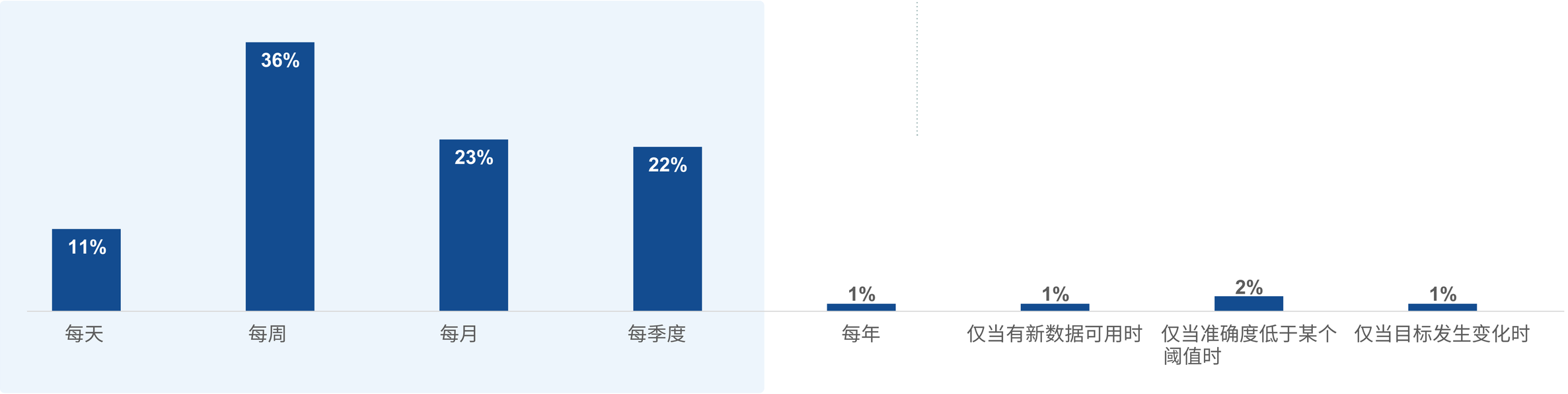
返回目录

> “一个明确定义的模型监控和维护策略，考虑到利益、成本和影响，对于做出正确的关于最佳重新培训计划的决策至关重要。”

在重新培训和维护之间取得平衡

由于47%的组织每周至少重新训练一次模型，了解频繁重新训练对组织的影响非常重要，包括资源紧张和效率低下，放大数据噪音和创建版本复杂性。 虽然基于数据漂移进行重新训练的变化很重要，但过度这样做可能会干扰运营，困惑用户，并妨碍对监控和伦理等关键部署方面的战略关注。 组织必须在重新训练频率和相关潜在弊端之间取得平衡。制定明确的模型监控和维护策略，考虑到利益、成本和影响，对于做出关于最佳重新训练计划的正确决策至关重要。

在生产中重新培训机器学习模型的频率。



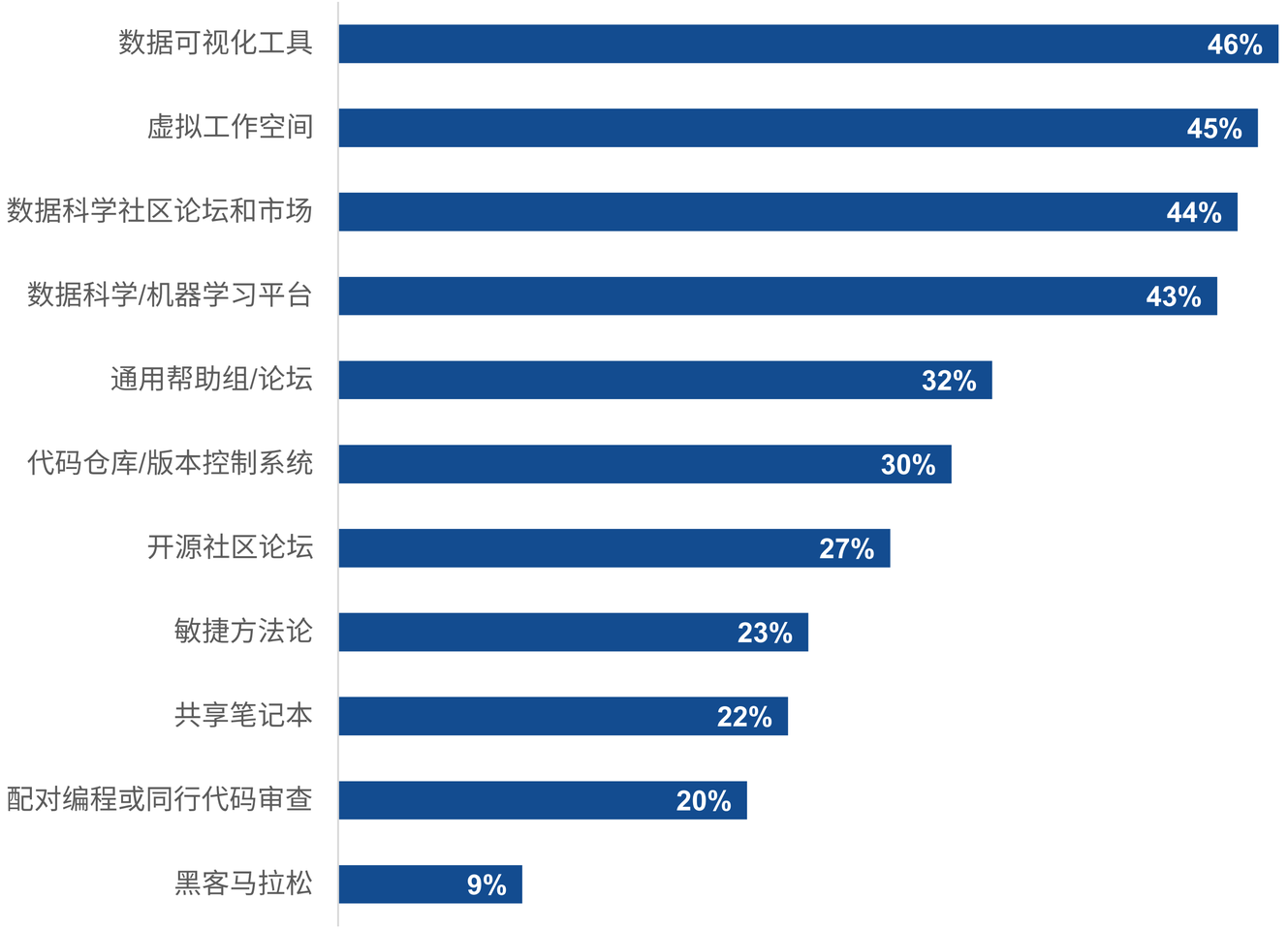
返回目录

数据科学和机器学习成为一个团队运动，供应商专注于使所有利益相关者能够参与其中

为协作数据科学的成功建立桥梁

各利益相关者和团队成员之间的合作对于成功的数据科学项目至关重要。
组织采用工具和方法来整合专业知识，促进建设性对话，策略完善和集体指导。 这种开放的沟通赋予不同角色塑造结果的能力，提高分析质量，并推动组织朝着变革性的洞察和决策迈进。

用于确保数据科学项目中各利益相关者和团队成员之间的协作的信息来源。



返回目录

在数据科学生命周期中映射利益相关者参与

非数据科学利益相关者在数据科学生命周期中扮演重要角色，影响从数据收集和预处理到模型部署和模型管理等各个阶段。 这是92%的受访者认为业务利益相关者参与数据科学项目并与数据科学团队合作的经验积极甚至非常积极的主要原因。 为非数据科学社区提供数据科学和机器学习解决方案为供应商带来重大机遇，因为无论组织的数据科学专业水平如何，它们都在数据科学领域不断前进。

涉及非数据科学专业人员（例如业务分析师）的机器学习模型构建领域。

44%
数据收集/供应

40%
数据预处理

39%
模型部署

38%
模型监控/维护

36%
模型训练

36%
模型评估

30%
模型选择

28%
逻辑构建

27%
用例/问题定义

返回目录

99%的受访者有动力提高
他们的数据科学和机器学习技能。

员工提高数据科学和机器学习技能的动力。

释放员工潜力

研究表明，99%的人受到改善数据科学和机器学习技能的动力，这种改善是内在动机和外在动机的结合。 职业晋升、认可和薪资增加的前景，以及对有意义地参与尖端项目的承诺，作为强大的外部动机。 这种有形奖励与知识好奇心的结合，在工作环境中创造了一个有趣的动态，员工受到激励投入时间（有时在工作之外）继续磨练他们的技能。

返回目录

Open for Innovation

KNIME

KNIME帮助每个人理解数据。

其免费且开源的KNIME分析平台使任何人（无论他们来自商业、技术还是数据背景）都能直观地处理数据，每天都能使用。KNIME商业中心是KNIME分析平台的商业补充，使用户能够在整个组织中进行数据科学的协作和洞察共享。 这些产品共同支持完整的数据科学生命周期，使各个分析准备水平的团队能够支持数据的操作化和构建可扩展的数据科学实践。

了解更多

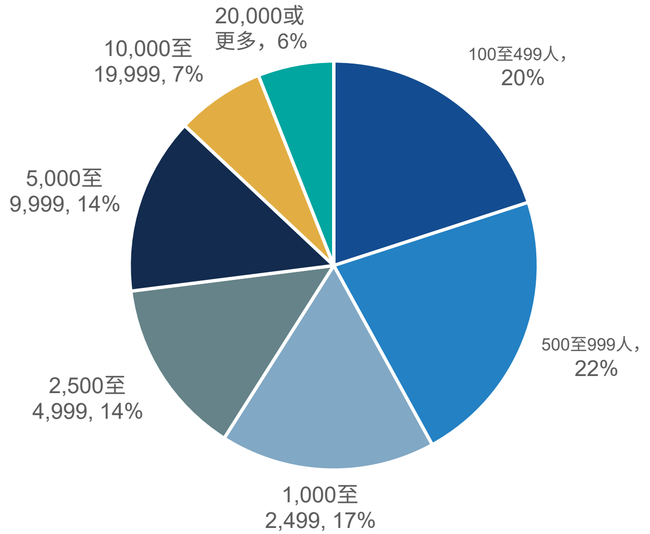
返回目录

研究方法和人口统计学

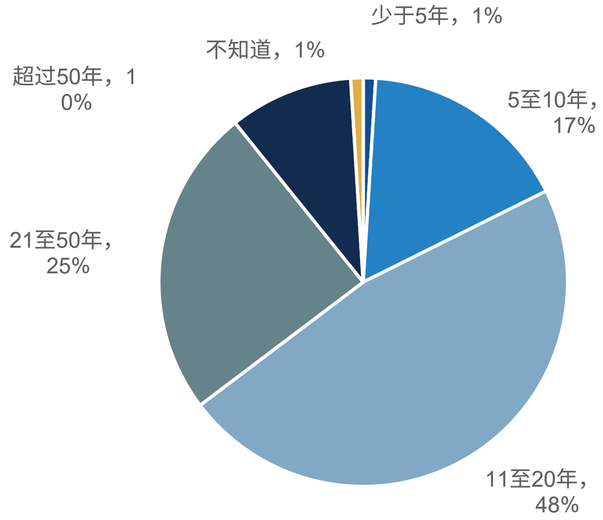
为了收集本报告的数据，ESG在2023年6月5日至2023年6月27日期间对来自北美地区（美国和加拿大）的私营和公共部门组织的数据专业人员进行了全面的在线调查。 为了有资格参加此调查，受访者需要涉及数据科学和机器学习技术和流程，包括可能负责制定战略、评估、购买、构建和管理这些技术。 所有受访者都获得了现金奖励和/或现金等价物作为完成调查的激励。

在筛选出不合格的受访者、删除重复回答并对剩下的完成回答（根据多个标准）进行数据完整性筛查后，我们最终得到了366名数据专业人员的样本。

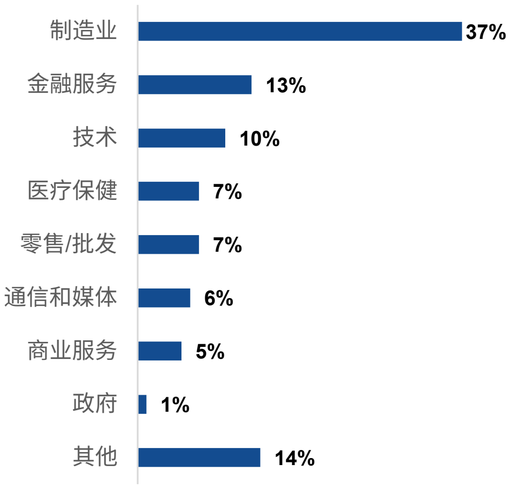
按员工人数分类的受访者



按公司年龄分类的受访者



按行业分类的受访者



返回目录