自主AI能力加速企业智能化转型

2022-2023爱分析
数据科学与机器学习平台应用实践报告

03. 2023

自主 AI 能力加速企业智能化转型

—2022-2023 爱分析·数据科学与机器学习平台应用实践报告

2023 年 3 月

报告编委

报告指导人

黄勇	爱分析	合伙人&首席分析师

报告执笔人

孟晨静	爱分析	分析师

外部专家（按姓氏拼音排序）

杜晨阳	力维智联	五维实验室主任
王哲	九章云极DataCanvas	雅图BU总经理

特别鸣谢（按拼音排序）

目录

CHAPTER

01

报告综述

1. 报告综述

随着数据体量的快速增长、算法迭代优化以及 CPU、GPU、DPU 等多种算力技术的发展，以大数据建模为核心的机器学习技术正被企业广泛应用到营销、广告、风控、生产等场景中。

机器学习涉及复杂的建模流程，如数据准备、特征工程、模型训练、模型部署、模型运营等，需要数据工程师、数据科学家、数据分析师、BI、软件工程师以及业务人员等多方协作。在企业传统的建模方式中，建模以项目制为主，建模周期长，协作困难，建模门槛高且严重依赖数学科学家。

然而，市场环境、消费者需求的快速变化推动企业向敏捷性组织转型，对业务决策时效性要求更加严格。对此，企业一方面需要提升建模效率以支持业务的持续更新、适应广泛的建模场景，另一方面也需要赋予一线业务人员建模能力，提升业务人员对市场的反应能力。传统建模方式难以满足企业快速决策需求。

数据科学与机器学习平台为企业提供了一个高效的解决方案。数据科学与机器学习平台整合数据接入、数据准备、特征工程、模型训练、模型部署、模型管理及模型运营等模型开发全流程，集成丰富的模型开发工具，不仅能有效提升模型开发效率，还能基于 AutoML 实现低门槛建模，满足业务人员的建模需求。数据科学与机器学习平台正成为企业数智化转型的必要基础设施。

不同行业的企业对数据科学与机器学习平台的需求侧重点不同。如对于具备专业建模人员的金融、医疗等行业，需要数据科学与机器学习平台兼顾专业建模人员和业务人员的建模需求；而对于普遍不具备专业建模人员的其他传统行业，如工业、消费、能源等，更需要业务人员可快速上手的低门槛建模系统。

本报告选取具有代表性的金融行业、工业行业的数据科学与机器学习平台解决方案为研究对象，围绕该解决方案在大中型企业的落地应用展开研究，重点分析两个行业中甲方对数据科学与机器学习平台的需求和解决方案。

CHAPTER

02

金融行业

数据科学与机器学习平台

2. 金融行业数据科学与机器学习平台

在领先的数字化转型进程、海量数据积累、充分的科技人才储备以及丰富的业务场景应用需求等驱动因素下，金融行业对数据科学与机器学习平台应用的渗透率明显高于其他传统行业。尤其在银行业，数据科学与机器学习平台的建设呈现出从全国性大型银行向地域性城商行覆盖的趋势。数据科学与机器学习平台作为人工智能基础设施正被纳入更多金融机构的数字化转型规划中。

以银行业为例，银行中的数据科学与机器学习平台的用户可分为两类人群：数据科学家和业务人员。其中数据科学家指具备专业建模能力的模型开发人员，负责模型的开发、算法的优化，是模型开发的核心人员。业务人员诸如营销、风控、产品研发等场景下的数据分析人员、BI 分析师。银行的 2C 属性使得更靠近 C 端消费者的业务人员对产品、服务的优化更敏感，也更具话语权，为实现银行的精细化运营，业务人员对敏捷地模型开发及应用的需求逐渐增强。两类人群对数据科学与机器学习平台的需求也不同。

图 1：数据科学家和业务人员对数据科学与机器学习平台的需求

数据科学家在进行机器学习建模时，主要面临以下挑战：

- 传统项目制建模方式导致计算资源无法共享：在金融机构传统的机器学习建模过程中，数据科学家各自以项目形式对业务场景进行建模，对于计算资源的调用以申请高性能 CPU 或 GPU 服务器为主，计算资源分配不均匀，算力不能高效利用。

- 传统建模方式下建模工具缺失：传统的开发工具简单，模型训练和模型部署都需要数据科学家手动实现，尤其模型部署过程中涉及模型转换、模型优化以及模型在业务平台运行的性能和稳定性等复杂的工程化落地能力，数据科学家实现模型部署较为困难。此外，由于缺乏数据、代码、模型的版本管理功能，建模过程中的数字资产无法共享、复用。
- 建模全过程多角色协同困难：由于模型开发过程会涉及到数据准备、模型训练、模型部署以及模型运维等多个环节，涉及数据工程师、数据科学家、软件数据分析师等多角色共同协作完成，存在反复沟通、协作流程不明确等问题，带来重复性工作。

业务人员对数据科学与机器学习平台的需求更偏向简单易上手的建模工具，需要屏蔽数据准备、模型训练、模型部署等环节的复杂性，实现一键建模，并能及时查看模型对业务决策分析的效果。

为同时满足数据科学家专业建模需求和业务人员低门槛的建模需求，最大化算法模型价值推动实现高效决策，金融行业的数据科学与机器学习平台解决方案应围绕以下要点展开。

图 2：金融行业数据科学与机器学习平台解决方案要点

- 统一资源管理：对模型开发需要的 CPU、GPU 资源进行整合，以容器化的方式对算力虚拟化，实现弹性扩容、性能加速、资源共享，避免资源浪费。
- 建立数据管道：模型训练过程依赖金融机构内外的高质量数据，且智能应用上线后，需持续对模型效果进行监控，持续输入新鲜的高质量数据集进行模型迭代，因此需要建立数据管道，包

括为金融机构接入多种数据源如关系型数据库、Hadoop 大数据平台，提供统一的存储、治理、管理服务，提供丰富的数据分析算子进行标注、检查、改进等数据预处理。

- 模型训练：兼容多种高性能训练和推理引擎框架，如 TensorFlow、Pytorch、MXNet 等。提供多种建模方式，包括自由度更高的 Notebook 建模、可视化建模、AutoML 建模，适用于金融机构不同建模人员使用。针对 Notebook 建模、可视化建模提供丰富的白盒算子，以供数据科学家进行优化或是建立模型训练工作流；AutoML 建模中则应具备数据自动处理、模型自动训练、模型自动选择等功能，使得业务人员只需提供原始数据集即可完成获得特定业务场景下的模型开发，开展智能应用。
- 模型部署和运维：提供一键部署功能，实现模型快速部署；提供模型监控功能，对模型漂移提供预警。
- 模型开发数字资产的沉淀：在模型开发过程中，针对数据接入、数据转换、特征工程、模型训练、模型部署等环节，提供数据、代码和模型等的版本管理，实现模型数字资产的沉淀和复用。

案例 1：AI 中心加速山西银行智能化转型，打造数据驱动型组织

山西银行是经中国银保监会批准，于 2021 年 4 月 28 日挂牌开业，以原大同银行、长治银行、晋城银行、晋中银行、阳泉市商业银行为基础，通过新设合并方式设立的省级法人城市商业银行，现有员工 7000 余名，拥有分行级机构 12 家，各类营业网点 387 个，遍布全省 10 个地市、23 个区、36 个县。

山西银行成立之初，在对原大同银行、长治银行、晋城银行、晋中银行、阳泉市商业银行科技系统整合的基础上，为建立一套全行的可持续“让数据用起来”的数据体系，于 2021 年启动数据中台项目群，推动包括数据开发平台、数据管控平台、数据服务平台和客户集市等功能实现。

建模方式不完善，亟待建模能力和建模系统全面升级

其中，为实现数据赋能业务需求，山西银行拟围绕以人工智能、大数据、云计算为代表的科技能力为基础搭建自动化联合建模平台，为建模人员提供样本导入、数据匹配、特征加工、模型训练及模型评估等一站式联合建模服务，并将联合建模平台作为数据开发平台的重要组成部分。山西银行对联合建模平台的需求主要体现在以下方面：

实现联合建模。山西银行中业务人员普遍不具备建模能力，而具备专业建模能力的科技人员对业务了解也不透彻，这导致科技人员在建模过程中需要与业务人员就具体需求、数据范围、数据质量、模型设计等方面进行反复沟通，耗费大量时间。山西银行亟需为业务人员实现自动建模功能，为科技人员提供一站式建模平台支撑，实现业务人员和科技人员联合建模，提升模型开发效率。

提升算力。AI 的算力强弱直接影响到 AI 模型训练的精度与推理结果。一方面，由于山西银行数据由 5 家银行数据合并而来，数据体量远超之前单个银行数据体量；另一方面，每个项目组都会各自申请计算资源，导致科技人员在进行模型训练过程中经常面临算力资源不足的问题，频繁出现内存溢出、开发工具重启等现象。此外，不同的业务场景需要的资源类型也不同，如机器学习模型常用 CPU 计算，深度学习模型倾向用 GPU 进行计算，因此如何提升建模的算力支持，且为科技人员屏蔽复杂的算力管理细节，专注于建模本身，是联合建模平台需要解决的主要问题之一。

实现数据、代码等模型数据资产共享及沉淀。山西银行技术人员在面向精准营销、智能风控、产品设计等不同业务需求时，优秀的数据集、代码、模型版本等成果不能及时共享，需要联合建模平台支持建模过程成果沉淀。

基于以上需求，山西银行将联合建模平台项目进行招投标，综合考量技术先进性、对业务场景的适应性、系统运行稳定性、系统安全性、系统可拓展性以及信创环境支持等因素，最终选择与九章云极 DataCanvas 合作。

北京九章云极科技有限公司（简称：九章云极 DataCanvas）成立于 2013 年，是中国数据智能基础软件领军者。公司专注数据智能基础软件的持续开发与建设，通过自主研发的一系列企业级 AI 应用所需的平台软件产品及解决方案，助力用户实现数智化升级。目前，九章云极 DataCanvas 机器学习平台业务涉及政府、金融、通信、制造、能源、交通、航空等十余个行业，客户覆盖多个行业头部和世界五百强企业。

基于 DataCanvas APS 机器学习平台，建设 AI 中心

在九章云极 DataCanvas 协助下，山西银行正式建设联合建模平台，基于九章云极成熟的 DataCanvas APS 机器学习平台建立“模型实验室”。该项目从 2021 年 11 月开始推进实施，历经近 9 个月的时间，于 2022 年 8 月初完成平台建设并进行线上试运行，之后于 2023 年 1 月正式在全行推广，针对全行范围的数据、模型需求正式开展工作。山西银行模型实验室面向科技人员和业务人员实现一站式模型开发，主要功能包括以下方面：

图 3：模型实验室功能架构图/示意图

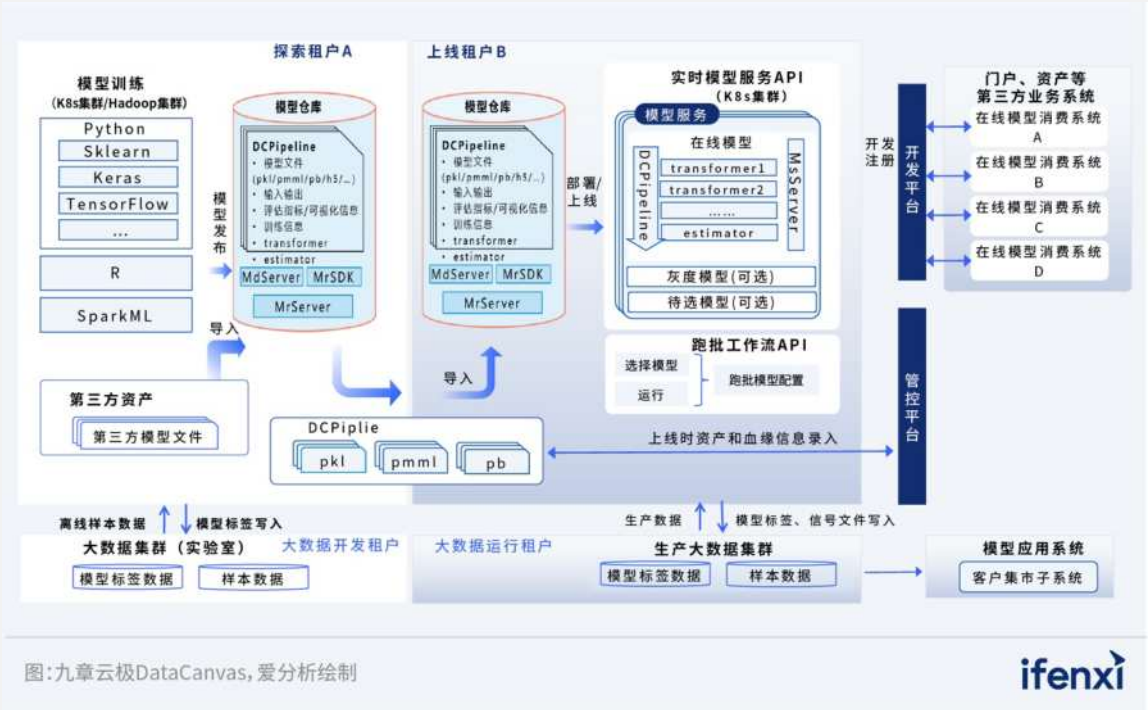


图:九章云极DataCanvas，爱分析绘制

ifenxi

1.异构多引擎融合架构

- 灵活计算环境支持：平台功能基于 Docker 实现容器化封装，底层计算资源支持 Kubernetes 集群、Hadoop 集群和 GPU 集群等多种模式，提供弹性可伸缩的 CPU 和 GPU 资源，支持大数据量的分析和训练，实现计算资源合理利用。
- 工作流混合编排：在异构多引擎融合架构下，平台算子封装支持多语言模式，允许在同一个工作流中调用不同开发语言算子，可以快速融合机器学习和深度学习的多引擎的训练和推理，支持工作流程嵌套，如在平台中支持编码、可视化、AutoML 三种建模方式，三种建模方式之间可相互调用，最大程度上提高建模流程的灵活性和模型资产的复用性。

2.简化数据准备，实现多源异构大数据分析

模型实验室支持多种数据连接器，山西银行可便捷获取包括本地数据、关系型数据库、Hadoop 大数据平台等在内的各类数据源，并且模型实验室支持支持异构多源数据的加工和混合处理，即在一个工作流中可以将多个异构数据源中的数据作为输入并调用平台上的多种数据分析算子进行处理。

3.开放性算法支持

- 集成了主流的开源机器学习算法库和深度学习框架，如 TensorFlow、Caffee、H2O 等，不同框架间可开展协同工作。
- 提供丰富的开箱即用“白盒”算法库，内置 100 多种算法模型，包括企业常用的统计分析、机器学习、深度学习算法，面向数据分析应用提供基础算法支持。“白盒”模式下，算子代码完全开放，支持客户对代码进行修改或开发，满足建模人员算子自定义、算子迭代需求。
- 建模人员可在集成 Web IDE 环境中，对算子进行开发。并基于容器技术对算子进行灵活封装、集成，形成算子模块并发布到算法库中。发布后的算子模块可被反复调用，提升新模型的开发效率。

4.提供三种编码方式，适应不同建模水平人员

- 代码建模：支持科技人员在 Web IDE 环境中通过 R、Python、Scala 等编程语言进行算法开发
- 可视化建模：模型实验室提供的算子模块覆盖模型生产全流程，包括数据准备、特征工程、模型训练、模型评估、模型对比、模型发布等，支持了解建模流程的科技人员通过图形化、拖拽式建模。
- AutoML 建模：针对不具备建模知识的业务人员，模型实验室提供低门槛 AutoML 技术，平台可自动完成包括算法选择、超参数优化、模型评估、模型选择及模型发布等系列过程，并生成面向生产系统的 REST API 调用服务。业务人员通过配置目标即可实现自动化建模。

5.模型全生命周期管理

对数据接入、数据转换、特征工程、建模可视化、模型仓库、模型生产化等建模全过程的数据、环境、代码、模型版本进行管理，实现数据、特征、模型的复用和迭代，沉淀数据资产。

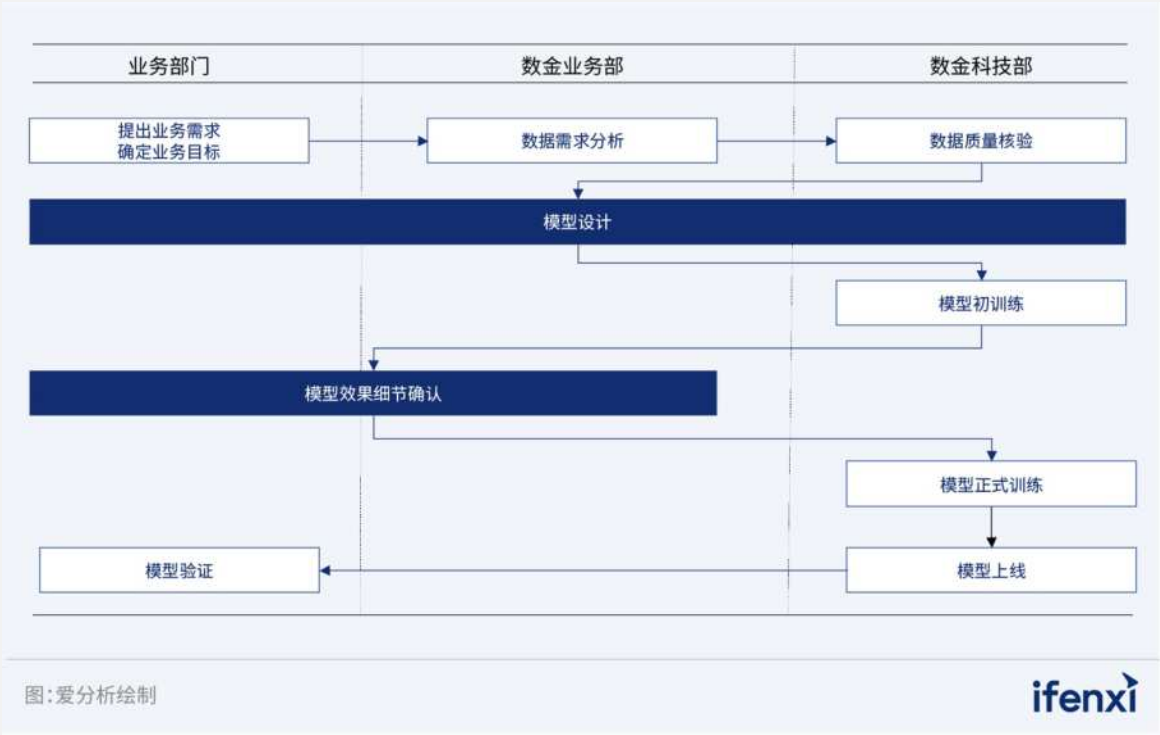
6.支持高性能的分布式训练

融合主流分布式计算框架如 Spark、TensorFlow、PyTorch、Dask 等，并预置丰富的分布式训练场景；深度学习分布式支持单机单卡、单机多卡、多机多卡训练，用户可以在复杂场景下快速高效完成模型训练。

以上是模型实验室的重要功能。

山西银行在搭建模型实验室的基础上，也在考虑如何改善模型开发流程让模型实验室发挥最大价值。由于模型开发流程包含业务需求分析、搜集数据、数据清洗、特征工程、模型训练、模型部署、模型运维等环节，涉及业务部门、IT 部门、算法开发人员等多个部门，为保证模型开发流程高效运转，在建设模型实验室基础上，山西银行制定了一套完善的模型开发协作机制，如下图所示。其中，业务部门提出业务需求并对模型最终效果进行确认。数金业务部承担与业务部门沟通的职责，包括业务需求确认、模型设计沟通、模型初训练的效果确认等。数金科技负责数据预处理、模型训练工作。

图 4：山西银行跨部门模型开发协作流程示意图



模型实验室大幅提升建模效率、有效降低建模成本

模型实验室作为山西银行的 AI 中心，利用先进的异构多引擎融合架构，适应业务人员和科技人员的不同建模需求，为智能应用建设生命周期提供完善的工具和支持，实现端到端一站式建模，有效解决算力瓶颈问题，大幅提升建模效率。

1.解决算力瓶颈问题

模型实验室基于异构多引擎融合架构，具有优秀的可扩展性，利用 Spark 分布式内存计算提供强大的计算能力，支持海量数据计算分析。此外，模型实验室能在模型开发的数据处理、模型训练等环

节提供资源自动推荐，用户也可对资源类型和配额进行调整，实现算力的高效利用。同时，模型实验室对使用者屏蔽了大数据技术组件的复杂性，使业务人员和科学人员能轻松获得大数据处理能力。

2.提升建模能力，提高建模效率

模型实验室提供端到端一站式建模全流程支持，能大幅提升山西银行在数据探索、预处理、特征工程、分析挖掘以及模型服务等环节的能力。另一方面，模型实验室为业务人员提供的 AutoML 建模和图形化建模方式，使业务人员能根据需求自主建模，基于模型效果再与科技人员沟通进行模型优化或调整，改进建模流程，大幅缩短建模时间，实现对业务需求的敏捷响应。

3.模型资产和建模方法论沉淀

建模过程中，包括数据集、数据清洗、特征工程、模型训练、模型上线等过程的代码、数据，以及建模的流程都能保留并提供下载，科技人员可以通过权限定义分享对象，从而实现人员协同、成果复用，沉淀模型资产、解决问题的方法论和流程。

4.有效实现成本控制：经统计，基于模型实验室，单个机器学习模型的建模成本缩减 60%，运维成本降低 30%。