



ifenxi

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

1. 报告综述	1
2. 金融行业数据科学与机器学习平台	3
3. 工业数据科学与机器学习平台	12
4. 结语	22
关于爱分析	23
研究咨询服务	24
法律声明	25

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

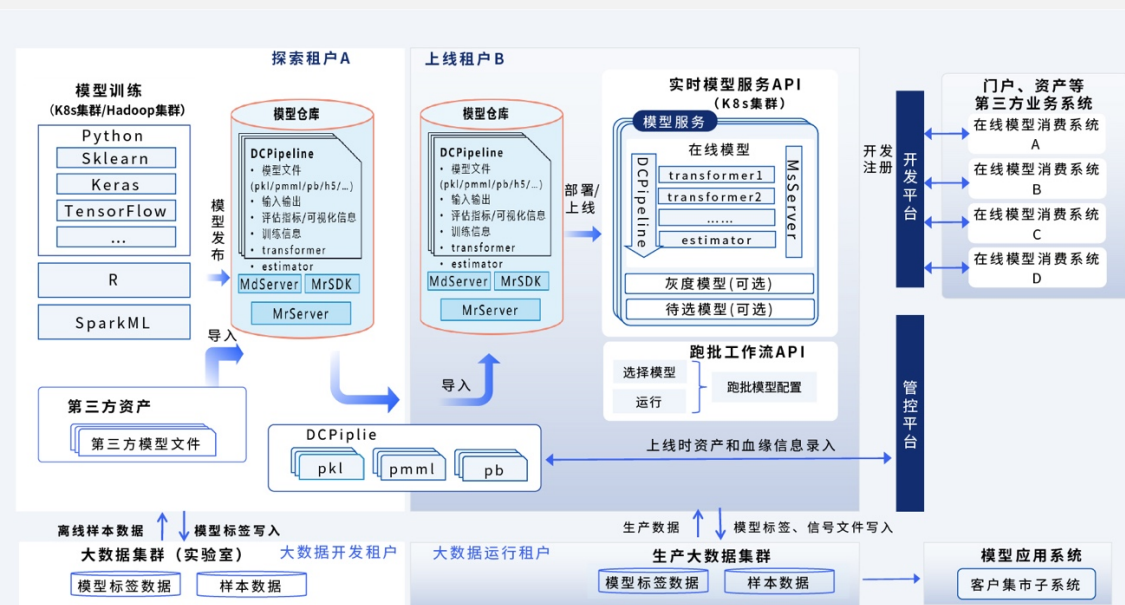


图:九章云极DataCanvas, 爱分析绘制

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、Caffe、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。

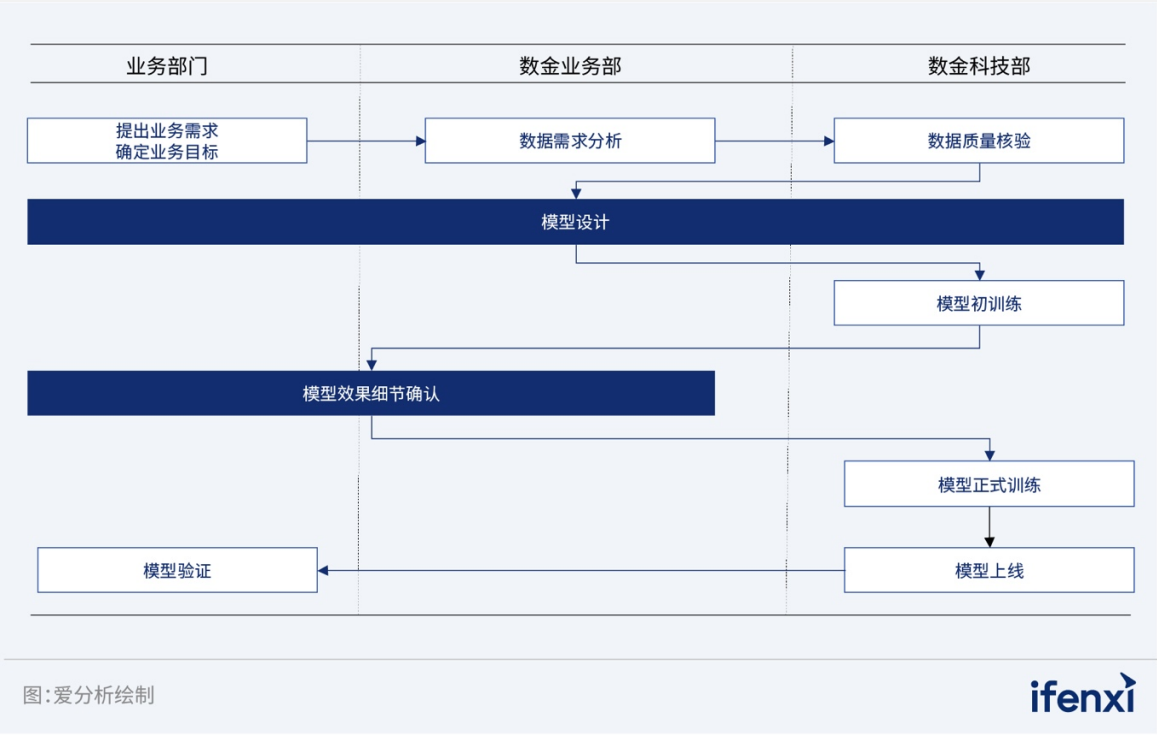
6.支持高性能的分布式训练

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%。

CHAPTER

03

工业

数据科学与机器学习平台

3. 工业数据科学与机器学习平台

工业互联网趋势下,以AI模型为核心的人工智能通过融合工业机理知识与专家经验,实现设计创新、生产优化、产品智能检测、智能运维等价值,正广泛应用在企业的设计研发、生产制造、运维、供应链管理、产品检测等场景中。人工智能已经成为工业企业提升生产效率、提高产品质量、降低人力成本、实现环境可持续发展的新型基础设施。

当前,工业企业中的数字化转型领先企业,在试点验证机器学习模型价值后,希望扩大智能应用的范畴,对特定场景进行智能化改造,如化工、石化等流程工业企业的生产场景下,生产装置的模拟与优化普遍基于传统的机理建模,以实现生产过程的工况分析和流程优化。但基于单个环节或是单个装置的机理模型收敛慢、研发周期长且模型可移植性差,难以实现对系统过程全流程的模拟。

机器学习平台通过融合机器学习建模和机理建模,不仅能简化模型复杂度,还能实现对生产过程中各环节、各化工装置实现建模开发,从而达到对生产全流程的精准控制和精准预测。工业企业对应用机器学习平台的难点/需求,具体表现在:

图 5: 工业企业应用数据科学与机器学习平台的难点/需求



- 企业缺失专业建模人才。工业企业的 IT 人才储备有限,企业不具备专业的机器学习建模人才。同时,企业的智能化进程需要为一线业务人员赋能模型开发、模型应用能力,因此机器学习平台必须低门槛、易上手。

- 实现知识资产沉淀和复用。工业下细分行业众多，不同细分行业的生产流程差异巨大。企业在日常经营中积累了丰富的场景知识和专家经验，需要将此类知识沉淀为数据科学与机器学习平台中的数据模板、特征工程模板、模型模板或是工作流程，实现专家经验知识沉淀。
- 提供个性化行业场景预训练模型支持。工业在生产工艺、产品检测等环节数据量有限，难以支撑需要大规模数据支撑的模型，需要合适的预训练模型实现小数据量下的模型开发。

为了解决工业企业在数据科学与机器学习平台应用上的难点，厂商对工业企业的解决方案应包括以下三方面内容：

图 6：工业企业数据科学与机器学习平台解决方案要点



1. 平台建设：搭建端到端一站式模型开发平台

- 建模方式：提供低门槛的建模支持，包括无代码 AutoML 建模和低代码图形化建模。其中 AutoML 应支持自动化的数据处理、特征工程、模型选择、超参数调优以及模型部署。
- 算法提供：提供丰富的数据预处理算法、机器学习算法、深度学习算法，供业务人员调用；提供业务场景适用的预训练模型，以保证业务场景下“小数据”的模型效果。
- 计算资源管理支持：支持算力资源的分布式管理和精细化管理，为模型训练提供自动化资源推荐，以及屏蔽大数据技术组件的复杂性，使业务人员能轻松获得大数据处理能力。

2. 平台部署：软硬件环境适配

工业企业的软硬件环境复杂，数据科学与机器学习建模平台内嵌在工业互联网平台中，面向多种业务系统，需要针对性的进行兼容性适配开发。因此，厂商软件开发人员需要与企业的工程师协同对认证系统、数据中台、业务系统进行定制化的开发和对接联调工作，以保证平台的顺利部署。

3. 平台运维：培训建模流程、技巧

在数据科学与机器学习平台建设完成后，厂商的数据分析师、数据科学家应提供培训教学，为企业的业务人员培训机器学习基础概念，同时通过实训方式让业务人员上机实操，结合业务场景案例，快速掌握平台建模能力。

案例 2：某石化企业通过工业大数据分析建模平台实现加氢裂化装置工艺优化，提升经营效率

某石化企业是一家集石油化工、煤化工、石化产品销售为一体，配套齐全的大型炼化一体化企业，原油加工能力 2200 万吨/年、芳烃生产能力 100 万吨/年，加工规模和技术水平位居国内炼油企业第一梯队。

近年来，该石化企业积极推进信息化工业化两化深度融合，基于工业互联网提出“智能炼厂”，在装置优化、计划优化、供应链优化等环节推进相关应用研究。目前，该企业工业互联网已广泛应用于工艺、设备、HSE、能源、经营管理、物流等领域。

与此同时，炼化企业为应对低油价与成品油市场寒冬，纷纷以生产优化为核心深入实施降本增效，进而对生产模型精度、模型开发效率以及模型应用广度等都提出更高要求。

而智能建模技术作为“智能炼厂”的核心技术，以炼油生产为“目标函数”，能通过快速创建智能化应用，为安全生产、降本增效提供智能决策，成为石化企业普遍关注的关键技术。加氢装置（包括加氢精制、加氢裂化）是石油炼化企业的关键环节，该石化企业希望通过敏捷建模实现炼化生产智能化。

既有建模方式门槛高、难应用

针对生产优化，该企业早已应用基于机理的传统优化建模软件如 Aspen、Petro-SIM、ProII，同时也在尝试基于机器学习算法的大数据建模技术，如将炼化机理知识与深度学习算法相结合，已经实现加氢装置原料油换热器结垢预测、加氢催化剂床层温度预测等场景的智能化应用，并取得了较好

效果。但两种建模路径都难以满足业务需求，使企业处于“有数据、无模型、有模型、难应用”的被动局面，具体痛点表现在：

1) 技术门槛高。基于机理的传统优化建模严重依赖专家经验以及国外厂商提供的优化求解软件，且建模软件模型收敛慢，调优复杂。而机器学习建模过程涉及数据准备、特征工程、模型构建、模型部署，也严重依赖具备专业建模能力的数据科学家。该企业缺乏懂建模也懂业务的专业复合型人才，同时高技术门槛阻碍了建模技术在企业炼化场景的广泛应用。

2) 建模周期长，建模成本高。传统优化建模和机器学习建模都存在着“大数据、小任务”，执行任务单一的特点。其建模过程中数据体量大、开发难度大、研发周期长，而开发的模型仅能适用于单个“小场景”。如果想覆盖炼化过程的更多场景就需要定制化的开发多个模型，导致两种建模方式均投入巨大，难以快速为企业带来效益提升。

3) 模型部署难。在该石化企业业务中，当机器学习算法模型训练完毕部署到生产环境中时，数据采集、预处理和边缘推理计算时均受智能设备影响，AI 模型的实际效果与智能设备的选型、接入、点位选择以及人力调试密切相关。然而该化工企业智能设备来源广泛，不同生产领域的设备型号各异，缺少统一的接入标准，且设备安装、接入和调试需要耗费大量人力，导致模型在不同生产领域部署时均需以项目课题形式实施，模型部署难也进一步制约机器学习建模在企业的推广应用。

建设一站式炼化工业大数据分析建模平台，支撑深度学习算法推广应用

鉴于机器学习建模在前期实验中的出色效果，该石化企业决定将深度学习算法进行推广应用，将“智能炼厂”推进“深度编码”阶段，通过先进的技术手段，如容器化、分布式计算、低代码等技术，实现对深度学习算法在炼化领域落地的助推。而实现深度学习算法的推广应用，需要大数据建模平台做支撑。该石化企业考虑到自建大数据建模平台成本投入大、技术难度较高，如分布式计算环境对硬件资源要求高、大数据组件的开发与应用难度大、模型部署对接各个系统复杂性较高等，于是决定与第三方厂商合作，综合考虑行业服务经验、核心技术能力、兼容性、成本等方面，最终选择与力维智联合作。

力维智联成立于 2005 年，是国家高新技术企业，凭借泛在数据接入与汇聚和大数据智能等 AIoT 技术和十余年 AIoT 系统集成交付能力，提供软硬一体的行业 AIoT 产品、解决方案与服务。同时，力维智联通过低代码机器学习开发平台提供 AI 开发工具软件和 AI 中台解决方案，该平台已服务于上

海世纪出版集团、南方电网科研院、中海油、航天科技、国家 5G 中高频器件创新中心等企业和公共服务平台。

在力维智联的协助下，该石化企业在既有工业互联网平台架构下，针对石化煤柴油加氢装置构建起一站式炼化工业大数据分析建模平台，集成数据预处理、主流机器学习算法、智能优化算法和模型自动训练方法，最终实现生产数据到模型应用的端到端建模平台。

图 7：工业大数据分析建模平台在工业互联网的位置



图 8：工业大数据分析建模平台架构示意图



1. 工业大数据分析建模平台架构

工业大数据分析建模平台处于石化工业互联网的工业 PaaS 层。架构上，工业大数据分析建模平台分为数据分析引擎和数据模型用户端。

1) 数据分析引擎

数据分析引擎分为底层架构和业务架构。底层架构由 Spark、MongoDB、Java Client、Webserver 等技术组成。业务架构包含数据读入、数据处理、特征工程、机器学习、深度学习、智能优化、AUTOML 等模块。数据分析引擎主要提供数据分析与挖掘、特征工程与模型训练功能，其中数据分析还包括数据输出及数据可视化。

2) 数据模型用户端

数据模型用户端底层架构由 Java Client、Azkaban、Redis、Java Springboot 等技术组成。业务架构通过数据视图——算子流构建——模型管理——模型工作流——模型服务——发布应用构建起来。数据模型用户端支持多种数据库接入，但不对接入数据进行存储，该石化企业另部署数据库进行对接。

数据模型用户端主要提供用户权限管理、资源的定义和管理、任务调度以及数据的呈现，其中资源包含数据视图、算子流、工作流和应用等。

2. 工业大数据分析建模平台功能

基于以上架构，工业大数据分析建模平台覆盖数据准备、统计分析、数据可视化、特征构造、模型构建和模型部署等智能应用开发全流程。具体实现了以下功能：

- 1) 提供丰富的算法：平台包含的算法主要分为两部分，一是通用数据科学挖掘算法，包括数据预处理和主流机器学习算法；二是优化算法，包括模型自训练算法和智能优化算法。一线炼化工程师可以根据自身需要设计算子流，对多种格式的数据数据进行预处理，或是根据业务场景进行建模。
- 2) 炼化数据与大数据融合：一方面，平台通过预置炼化过程“工艺、质量”等领域专题数据集固化先验知识，并对工艺特征工程进行迁移，为大数据算法落地提供特征模板与数据基础，将工程师的专家知识固化到平台；另一方面，平台预置丰富机器学习、深度学习算子，可为不同炼化过程大数据建模场景提供算法与数据的适配融合。
- 3) 实现低门槛建模：一方面，基于应用引擎、算子流和工作流的编排功能，工业大数据分析建模平台为炼化工程师提供无代码的图形化“拖拉拽”建模支持，通过连接任意算子形成算子流，可便捷实现数据预处理、模型设计、训练和部署。另一方面，平台支持无代码的模型自训练（AutoML），包括自动数据准备，自动机器学习，自动模型调参，大大降低炼化工程师的建模门槛。
- 4) 提供智能应用开发：模型部署后，平台提供 API 服务的方式供其他业务系统调用，同时也支持借助 BI+AI 图标和控制组件开发动态可交互的可视化应用。
- 5) 资源分布式、项目制管理：通过集群部署，平台实现了模型训练算力与推理算力资源的分布式管理，以项目为单位，对数据资源和硬件资源进行精细化管理控制。
- 6) 软硬件兼容适配：力维智联在平台部署时使用了容器化技术，对国产化软硬件进行兼容性适配开发，保证平台顺利部署、稳定运行。针对数据接口接入，力维智联通过适配该炼化企业自有的工业互联网平台数据微服务接口，完成了炼化数据接入，保证数据流转畅通。另外，平台还额外提供了 ETL 能力，对数据进行治理与融合，实现算法建模前对数据进行预处理。
- 7) 完备的流程管理：平台提供从数据接入、数据分析处理、数据展示、特征工程、模型管理、模型部署以及智能应用上线一站式实现和管理。

工业大数据分析建模平台显著提高建模效率、转变模型开发方式

该石化企业通过大数据低代码建模平台，解决了炼油化工企业“有数据、无模型；有模型、难应用”、“建模成本高、建模周期长”的问题。通过实践证明在预测误差、利润产品收率、建模周期等方面均得到显著改善。

1) 模型性能提升，预测误差降低

基于大数据建模平台开发的“多通道多尺度卷积神经网络（MCMSCNN）”模型在加氢催化剂床层温度预测、氢气耗量预测方面误差<3%，提升了炼化工程师对核心工况参数的监控把握，大幅降低氢气、燃料气等生产成本。

基于低代码平台开发的“大数据预训练+工艺数据微调”迁移算法解决了因LIMS（实验室信息管理系统）化验分析数据量不足而造成的产品性质预测模型性能不佳的问题，实现了加氢裂化产品性质的实时软测量，利润产品收率提高约0.8%。

2) 建模门槛降低，建模周期大幅缩短

数据智能建模平台实现了生产装置海量数据的有效利用，从数据采集到模型部署，由2周大幅缩短为1天。其中建模环节，因平台支持通过建立与炼油装置有关的工艺和质量的专题数据集来固化先验知识，迁移建模特征，为大数据算法落地提供特征模板与数据基础，使建模时间由原来的一周大幅缩短为8小时，建模效率大幅提高。

此外，该石化企业内的模型开发工作也实现了“去中心化”，从原有信息中心的个别工程师在实验室开发，转变为一线工程师结合生产问题广泛开发，有效赋能生产。

3) 模型资产实现复用

由炼化工程师自主研发的10余个机器学习模型，日调用量超1000余次。

厂商选型和企业科研意识是大数据低代码建模平台落地关键

该石化企业通过建设低代码工业大数据分析建模平台实现加氢装置多环境的优化，取得良好效果，其成功经验可以复用到炼化行业生产工艺环节、其他流程行业企业，或是供已建设工业物联网的企业借鉴，主要包括以下两点：

1) 厂商选型

石化企业的生产流程具有显著行业特点，生产环境复杂，对产品的稳定性、可靠性要求苛刻，且石化企业普遍不具备建模能力，建模平台完成后需提供完善的售后服务。因此厂商应具备以下能力。

- 厂商应具备工业、制造业或是能源行业的服务经验，具备一定的行业 Know-how 积累，理解行业工业互联网要求、了解企业业务流程。
- 厂商产品须为自主研发，能针对企业个性化需求进行定制化开发。
- 厂商产品及技术应兼容国产化软硬件环境，能够在国产 CPU、AI 加速芯片以及操作系统上稳定运行。
- 厂商应能提供数据分析师、数据科学家为企业提供培训教学，让一线工程师快速掌握基础概念，结合业务案例，通过实训的方式让工程师掌握平台使用技能。

2) 企业自身对数字化、对科研的重视

一方面，企业需要有强烈的科研意识，能够紧跟新技术发展，勇于在自身业务领域进行创新；另一方面，企业管理层对工业互联网、对数字化建设的重视也是机器学习建模平台能顺利落地的保障。

CHAPTER 04

结语

关于爱分析

爱分析是一家专注数字化市场的研究咨询机构，成立于中国数字化兴起之时，致力于成为决策者最值得信赖的数字化智囊。

凭借对新兴技术和应用的系统研究，对行业和场景的深刻洞见，爱分析为数字化大潮中的企业用户、厂商和投资机构，提供专业、客观、可靠的第三方研究与咨询服务，助力决策者洞察数字化趋势，拥抱数字化机会，引领中国企业数字化转型升级。

研究咨询服务

技术研究

新兴技术研究，厂商能力调研，助力数字化最优决策

商业研究

基于研究、数据和案例调研积累，辅助业务可靠落地

客户洞察

企业用户需求及实践调研，辅助制定业务与市场策略

品牌&营销

权威背书，树立行业地位；教育市场，精准触达客户

行业研究

行业数字化趋势与实践研判，辅助业务与战略决策

投资研究

成熟方法论，一手数据，助力研判机会、稳健投资

法律声明

此报告为爱分析制作，报告中文字、图片、表格著作权为爱分析所有，部分文字、图片、表格采集于公开信息，著作权为原著者所有。未经爱分析事先书面明文批准，任何组织和个人不得更改或以其他方式传送、复印或派发此报告的材料、内容及其复印本予任何其它人。

此报告所载资料的来源及观点的出处皆被爱分析认为可靠，但爱分析不能担保其准确性或完整性，报告中的信息或所表达观点不构成投资建议，报告内容仅供参考。爱分析不对因使用此报告的材料而引致的损失而负上任何责任，除非法律法规有明确规定。客户并不能仅依靠此报告而取代行使独立判断。

北京爱分析科技有限公司 2023 版权所有。保留一切权利。



咨询/合作

微 信: ifenxi888

网 址: www.ifenxi.com

地 址: 北京市朝阳区酒仙桥南路2号院东风kaso4层406



如欲了解更多爱分析精彩洞见
请关注我们的微信公众号

©北京爱分析科技有限公司2023版权所有



ifenxi
专注数字化