NAIE
V200R021C10
训练服务

文档版本 01

发布日期 2020-08-30

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

目 录

1 文档导读

NAIE模型训练服务提供了产品介绍、快速入门、用户指南、常见问题和产品术语手册，帮助用户快速熟悉和使用NAIE模型训练平台，进行模型训练和模型管理。

表 1-1 文档导读

手册	概述
《产品介绍》	本文档详细阐述了NAIE模型训练服务的定位、优势、功能、架构与适用场景等。
《快速入门》	本文档以硬盘故障检测的模型训练为例，介绍NAIE训练平台使用的全流程，包括数据集、特征工程、模型训练、模型管理和模型验证，使开发者快速熟悉NAIE训练平台。
《用户指南》	本文档包含了使用NAIE训练平台前的准备工作和如何使用NAIE训练平台导入数据、特征操作、模型训练、模型打包与模型验证的操作指导。
《学件开发指南》	本文档介绍KPI异常检测学件使用的全流程，包括数据集、模型训练、模型管理和发布在线推理服务。
《常见问题》	本文档收集了用户日常使用NAIE训练平台遇到的问题，并给予解答，有助于快速解答用户问题。
《产品术语》	本文档详细阐释了NAIE模型训练服务相关的产品术语。

2 产品介绍

2.1 什么是模型训练服务

模型训练服务为开发者提供电信领域一站式模型开发服务，从数据预处理，到特征提取、模型训练、模型验证、在线推理，本服务为开发者提供开发环境、模拟验证环境，API和一系列开发工具，帮助开发者快速高效开发电信领域模型。

2.2 产品优势

电信经验嵌入降低模型开发门槛

- 集成50+电信领域AI算子&项目模板提升训练效率，降低AI开发门槛，让开发者快速完成模型开发和训练
- AutoML自动完成特征选择、超参选择及算法选择，提升模型开发效率
- 高效开发工具JupyterLab和WebIDE：交互式编码体验、0编码数据探索及云端编码及调试

联邦学习&重训练，保障模型应用效果

- 支持联邦学习，模型可以采用多地数据进行联合训练，提升样本多样性，提升模型效果
- 支持迁移学习，只需少量数据即可完成非首站点模型训练，提升模型泛化能力
- 模型自动重训练，持续优化模型效果，解决老化劣化问题

预置多种高价值通信增值服务，缩短模型交付周期

- 无需AI技能，支持模型自动生成，业务人员快速使用
- 多种通信增值服务开箱即用，快速支撑电信领域AI应用

支持 3 种部署模式

- 公有云部署：数据允许出局，面向用户包括：中小T、合作伙伴、华为内部研发。
- 合营云部署：数据不出局，面向用户为有合营云的大T。

- 华为云Stack部署：数据不出局，面向用户为无合营云的大T。

2.3 功能介绍

数据集

导入模型训练使用的数据集，提供最大值、最小值、均值、方差、可视化数据分析能力对数据质量进行评估分析。

特征工程

特征工程是模型训练的必要过程，可以实现数据集的特征组合、筛选和转换，最大限度的从数据集中提取关键特征，供模型训练使用。当前已经支持电信领域的业务对象，如基站、交换机、路由器等设备的特征处理能力，辅助发现关键特征，提升模型训练效果 。

模型训练

提供线上的简易编辑器环境和在线VS code IDE编程工具，支持开发者在线切换，协同开发模型，支持华为自研AI框架MindSpore和业界多种主流AI计算框架Tensorflow、Spark MLlib、MXNet、PyTorch等，可并行提交多个模型训练任务，支持集成学习，提供GPU、CPU计算资源供开发者选择。

模型管理

训练模型的开发和调优往往需要大量的迭代和调试，数据集的变化、训练算法或者超参的变化都可能会影响模型的质量。用户可将训练完成的优质模型打包到模型管理中，进行统一管理。支持如下功能：

- 新建模型包（主要应用于多个模型打包成一个模型包的场景）
- 删除模型、下载模型
- 通过在线VS code IDE编程工具编辑模型和模型相关的数据处理等能力
- 将模型上架至AI市场
- 将模型发布成在线推理服务、更新发布在线推理服务
- 模型包完整性校验
- 创建联邦学习实例

模型验证

模型验证是基于新的数据集或超参，对训练平台已打包的模型进行验证，根据验证报告判断当前模型的优劣。

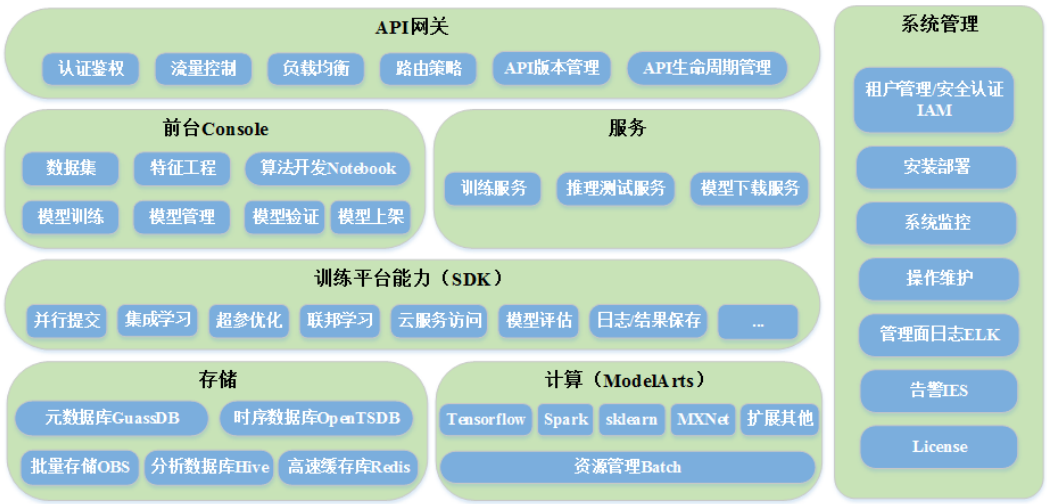
云端推理框架

提供云端的模型运行框架环境，支持将AI模型快速发布成云上的实时推理服务、对外提供可调用的服务API，帮助用户高效低成本地完成模型的部署、验证与发布。

2.4 产品架构

训练平台的产品架构图，如图2-1所示。

图 2-1 产品架构



训练平台产品架构图说明，如表2-1所示。

表 2-1 产品架构说明

功能模块	描述
API网关	训练平台API接口能力。
前台Console	训练平台在线IDE能力。
服务	训练平台对外提供的服务。
训练平台能力	训练平台提供的SDK能力，支持扩展。
存储	训练平台的存储应用。
计算（ModelArts）	训练平台集成的华为云服务的ModelArts能力。
系统管理	训练平台的系统管理能力。

2.5 适用场景

在面向网络资源效率提升、能源效率提升、运维效率提升、用户体验提升等方面的业务场景时，模型训练服务为网络通信领域，包括无线、固定网络、核心网、数据中心

四个领域的相关人员，提供AI集成开发环境，使开发者可以训练出一个模型，并提供模型验证功能。

华为产品线用户

开发AI算法，利用数据服务里的数据，生成模型，提供给运营商使用。

运营商用户

- 三产公司基于自己的数据，使用训练服务开发AI算法，生成模型供自己使用。
- 从AI应用市场订购并下载模型，部署至推理框架后，进行推理应用。
- 使用模型训练服务打包的模型，发布成在线推理服务，进行在线实时验证。

高校科研用户

开发AI算法，利用数据服务里的数据，生成模型，做相关AI算法研究、论文发表。

生态合作伙伴

开发AI算法，利用数据服务里的数据，生成模型，发布到AI应用市场，供用户订购。

2.6 基本概念

AI 引擎

可支持用户进行机器学习、深度学习、模型训练作业开发的框架，如Tensorflow、Spark MLlib、MXNet、PyTorch等。

数据集

某业务下具有相同数据格式的数据逻辑集合。

数据准备

数据集中导入的数据实例，可能存在空值、冗余、数据不足等情况，或者用户需要进行数据连接、数据联合、数据修复等操作。

在旧版体验式开发模式下，数据准备包含的功能有：数据修复、数据过滤、数据联合、数据连接、数据去噪。对应JupyterLab交互式开发模式界面“算子 > 数据处理”菜单下面的部分数据处理项。

特征操作

特征操作主要是对特征的样本数据值进行修改，也可以重命名、删除、筛选特征列。

在旧版体验式开发模式下，训练平台支持的特征操作有重命名、归一化、数值化、标准化、特征离散化、One-hot编码、数据变换、删除列、选择特征、卡方检验、信息熵、新增特征、PCA。对应JupyterLab交互式开发模式界面“算子 > 数据处理”菜单下面的部分数据处理项。

模型包

训练模型的原始包，包括模型文件。可以基于模型包创建模型验证服务、训练服务。可以上架至应用市场，支持用户订购后，下载到推理框架中使用。

2.7 与其他服务的关系

与 ModelArts 服务的关系

NAIE平台使用，华为公有云系统提供的ModelArts服务，实现数据预处理、大规模分布式模型训练能力。

与统一身份认证服务 IAM 的关系

NAIE平台使用，华为公有云系统提供的统一身份认证服务（Identity and Access Management，简称IAM），实现统一的身份认证和权限管理服务。

与 API 网关的关系

NAIE平台必须对接到华为公有云系统提供的统一API网关，此API网关为用户提供统一的入口调用NAIE云服务的API。NAIE云服务开放给租户的API，必须在API网关上注册通过后再发布。

与对象存储服务的关系

NAIE平台使用对象存储服务（Object Storage Service，简称OBS）存储数据和模型的备份和快照，实现安全、高可靠和低成本的存储需求。

与云容器引擎的关系

NAIE平台使用云容器引擎（Cloud Container Engine，简称CCE）部署模型为在线服务。支持服务的高并发和弹性伸缩需求。

2.8 计费说明

计费项

模型训练服务按照用户选择的实例规格和使用时长计费。计费项包括模型训练环境和云上推理服务，如表2-2所示。

表 2-2 计费项

计费项	计费说明
模型训练服务	模型训练服务根据CPU和GPU的规格和使用时长进行计费，不使用则不产生费用。 当训练服务开始启动以后，实例处于Running状态时，开始计费。请及时停止不需要的实例，以免产生不必要的费用。

计费项	计费说明
云上推理	云上推理服务根据CPU和GPU的规格和使用时长进行计费，不使用则不产生费用。 当模型一旦部署在云上推理服务中，并启动运行，实例处于Running状态时，开始计费。请及时停止不需要的实例，以免产生不必要的费用。

计费模式

按需计费模式，即按处于Runing状态的实例规格和实际使用时长计费。

- 计费公式为：单位价格 * 实例数量 * 使用时长，截取到“分”扣费。
- 按需付费模式下，如果价格计算器上的金额遇到小数点，保留小数点后两位，第三位四舍五入。如果遇到四舍五入后不足0.01元，则按0.01元展示。
- 模型训练服务会使用对象存储服务（OBS）。

变更配置

订购模型训练服务不会产生费用，运行实例才会产生费用，因此不涉及服务变更配置。用户可以根据实际情况，选择相应规格的实例运行。

续费

用户可以根据实际使用情况，及时充值。保证可以正常使用模型训练服务。

到期与欠费

如果没有按时续费，云平台会提供一定的宽限期和保留期。宽限期和保留期的市场由客户等级而定，详情请参见“**宽限期保留期**”。

保留期满仍未充值，则系统会清空资源。

2.9 如何访问模型训练服务

步骤1 在用户终端PC上打开浏览器，在地址栏中输入“**https://console.huaweicloud.com/naie/**”，进入AI市场。

步骤2 单击界面右上角的“登录”，进入登录界面。

步骤3 选择“IAM用户登录”方式，输入租户名、用户名和密码。

用户也可以直接通过账号登录。首次登录后请及时修改密码，并定期修改密码。

步骤4 单击“登录”，进入AI市场。

步骤5 依次选择“AI服务 > 模型与训练服务 > 模型训练 > 模型训练服务”，进入模型训练服务介绍页面。

步骤6 单击“进入服务”，进入模型训练服务页面。

----结束

2.10 修订记录

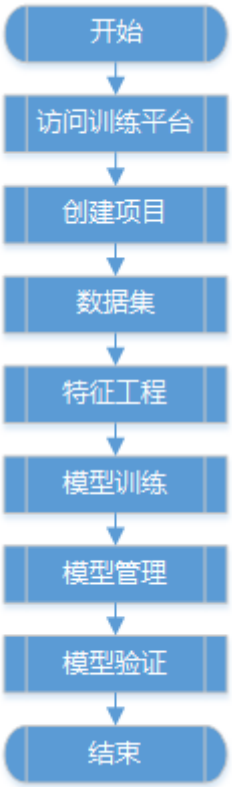
发布日期	修订记录
2020-08-30	根据最新的训练平台，更新如下章节内容： • **产品优势** • **功能介绍** • **基本概念**
2020-06-30	新增“计费说明”章节。
2019-12-30	服务功能优化，资料全量刷新。
2019-04-30	第一次正式发布。

3 快速入门

3.1 操作流程

模型训练服务操作流程如**操作流程图**所示。

图 3-1 操作流程图



3.2 前提条件

- 已经注册华为云账号。

- 已经注册NAIE平台的管理员租户和IAM用户。
- 已经订购过NAIE模型训练服务。

3.3 订购模型训练服务

步骤1 在用户终端PC上打开浏览器，在地址栏中输入“**https://console.huaweicloud.com/naie/**”，进入AI市场。

用户首次访问AI市场，会进入“访问授权”界面，单击“授权并继续”即可。

步骤2 单击界面右上角的“登录”，进入登录界面。

步骤3 输入租户名和密码，单击“登录”，进入AI市场。

首次登录后请及时修改密码，并定期修改密码。

步骤4 依次选择“AI服务 > 模型与训练服务 > 模型训练 > 模型训练服务”，进入模型训练服务介绍页面。

步骤5 单击“我要购买”，进入如**图3-2**所示的界面。

区域：为用户提供服务的华为云Region。请选择“华北-北京四”。

用户可以单击“了解计费详情”，详细了解训练服务提供的资源、规格和相应的价格信息。同时，用户在使用具体资源时，训练服务会在界面给出醒目的计费提示。

图 3-2 订购训练服务

步骤6 单击“立即使用”，服务订购完成。

----结束

3.4 访问模型训练服务

步骤1 在用户终端PC上打开浏览器，在地址栏中输入“**https://console.huaweicloud.com/naie/**”，进入AI市场。

步骤2 单击界面右上角的“登录”，进入登录界面。

步骤3 选择“IAM用户登录”方式，输入租户名、用户名和密码。

用户也可以直接通过账号登录。首次登录后请及时修改密码，并定期修改密码。

步骤4 单击“登录”，进入AI市场。

步骤5 依次选择“AI服务 > 模型与训练服务 > 模型训练 > 模型训练服务”，进入模型训练服务介绍页面。

步骤6 单击“进入服务”，进入模型训练服务页面。

----结束

3.5 创建项目

步骤1 在训练平台首页，单击“创建项目”上方的“+”按钮，创建硬盘故障检测工程。

参数设置如**图3-3**所示。

参数含义如下所示：

- 模板：复用已有的电信经验创建项目。项目创建后，项目中会预置有相关的数据集、特征处理操作、模型训练算法以及模型验证算法。
- 是否公开：创建用户的时候可以设置所属的用户组，如果勾选，则展示“公开至组”参数。
- 公开至组：默认展示当前用户所属的所有用户组，如果勾选用户所属的用户组，则被勾选用户组下的所有用户均可以查看当前用户创建的项目。
- 图标：项目的图标。用户可以自行从本地上传图片。

图 3-3 创建项目

步骤2 单击“创建”，完成工程的创建。

进入项目概览页面。

说明

用户当前的操作正处在项目概览页面、数据集页面、特征工程页面、模型训练页面、模型管理页面或模型验证页面时，如果需要回到训练平台首页，请单击界面左上角品牌右侧的HOME标识内容，从下拉菜单中选择“模型训练服务”。

----结束

3.6 数据集

针对硬盘故障检测，一共准备了四份数据集，分别如下所示：

- HardDisk-Detect_Train_Good.csv：无故障硬盘训练数据

- HardDisk-Detect_Train_Fail.csv：故障硬盘训练数据
- HardDisk-Detect_Test_Good.csv：无故障硬盘测试数据
- HardDisk-Detect_Test_Fail.csv：故障硬盘测试数据

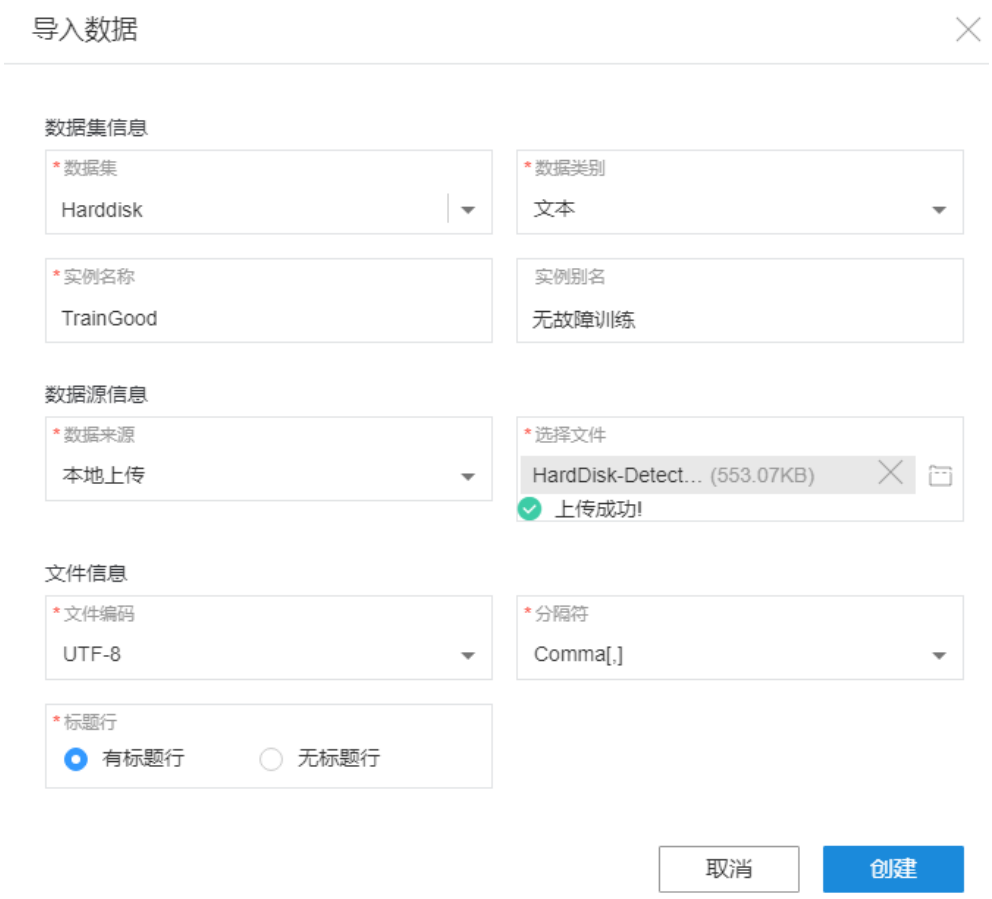
步骤1 在“项目概览”部分，单击“数据集”下的“创建”。

进入“数据集”界面，界面自动弹出如图3-4所示的对话框。

创建无故障训练数据集，参数含义如下所示：

- 数据集：默认为“Default”，支持自定义输入，例如：Harddisk。单击“创建”后，自动在左侧导航中，新增Harddisk节点。
- 实例名称：支持自定义。此处设置为：“TrainGood”。
- 实例别名：支持自定义。此处设置为：“无故障训练”。方便识别数据。
- 数据来源：下拉框中有三个选项，一是“本地上传”，即从本地上传数据文件，文件会自动上传至OBS租户空间中。二是“数据目录”，在用户已经订阅数据集的情况下，订阅并选择数据集文件，导入至训练平台。三是“样例数据”，即训练平台预置的样例数据。

图 3-4 导入数据



步骤2 单击“创建”，数据文件自动上传至用户OBS租户空间中。

步骤3 单击数据所在行，对应“操作”列的图标。

进入数据操作界面，如图3-5所示。

图 3-5 数据操作界面

步骤4 单击导入状态旁的“元数据”。

进入数据分析界面。

步骤5 设置引擎和规格，单击界面右下角的“分析数据”。

数据分析完成后，数据详情信息如图3-6所示。

图 3-6 数据详情

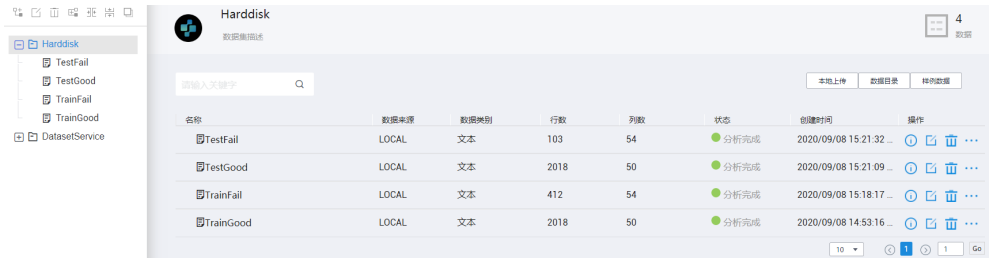
步骤6 单击数据预览界面右上方的×，返回数据操作界面。

步骤7 单击左侧导航中的数据集节点“Harddisk”，回到数据集首页。

步骤8 请参考步骤1~步骤7，单击界面右上角的“本地上传”，分别创建故障硬盘训练、无故障硬盘测试和故障硬盘测试数据集并完成数据分析。

创建完成后，界面可以看到四份数据，如硬盘故障检测所示。

图 3-7 硬盘故障检测



----结束

3.7 特征工程

步骤1 单击无故障硬盘训练数据集所在行对应“操作”列的...，在展开的下拉菜单中单击  图标。

弹出“特征处理”页面，如图3-8所示。

参数说明如下所示：

- 开发模式：特征工程开发环境。请选择“Jupyterlab交互式开发”。
- 规格：资源配置信息，请按需求选择，如选择“2核|8G”。
- 实例：无环境实例时，请从下拉框中选择“新建一个新环境”。

特征工程中的特征操作详情，请参考模型训练服务的《用户指南》中的“特征工程”章节内容。

图 3-8 特征处理

步骤2 单击“创建”。

弹出特征工程界面，待系统完成创建后，新建特征工程的“环境信息”旁显示特征工程状态为“运行中”。

步骤3 单击新建特征工程“操作”列中的  。

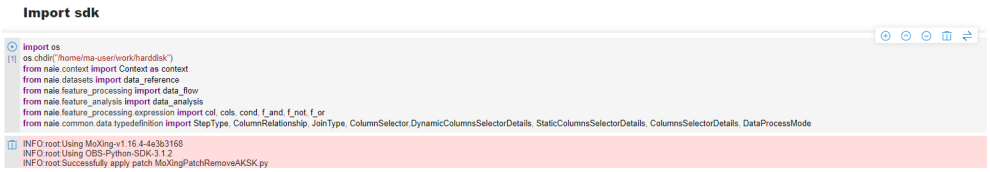
进入特征工程JupyterLab环境编辑界面，默认打开与特征工程同名，后缀为“ipynb”的特征工程操作主文件。

注意

在进行数据处理操作前，请先运行“Import sdk”代码块，否则会导致“选择数据”出错。

步骤4 在右侧特征工程操作主文件上，单击“Import sdk”代码块的运行按钮，运行导入SDK代码块，如图3-9所示。

图 3-9 运行“Import sdk”



步骤5 单击展开特征工程JupyterLab环境编辑界面右侧的“算子”菜单，选择“数据处理”页签，在“数据集”栏目下单击“选择数据”，或者单击“Import sdk”代码块下方代码块中的“选择数据”。

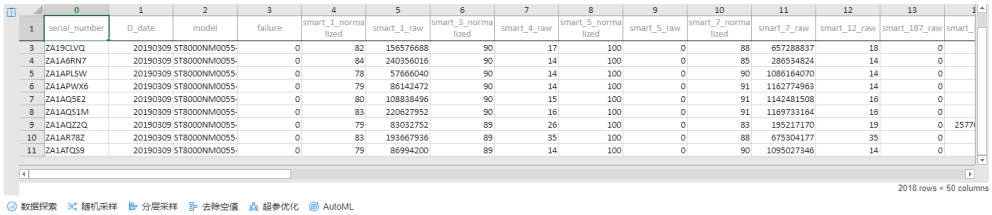
步骤6 在“选择数据”代码块中设置数据集及数据集实例，单击运行代码块。如图3-10所示。

- “数据集”设置为“**数据集**”中**步骤1**步骤内设置的数据集。
- “数据集实例”设置为“**数据集**”中**步骤1**步骤内导入的无故障硬盘训练实例。

图 3-10 选择数据

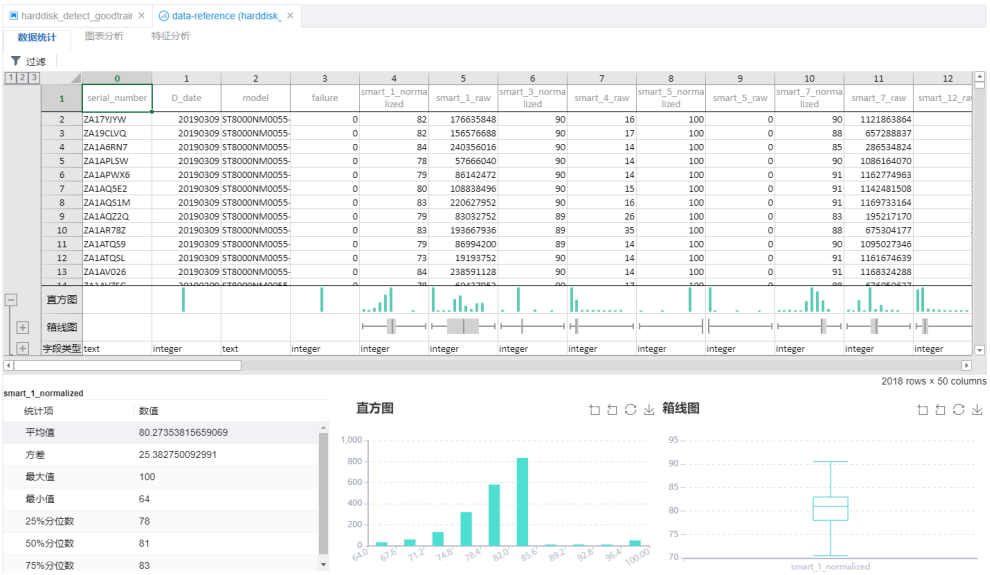
运行成功后，“选择数据”代码块下方展示特征数据表，如图3-11所示。

图 3-11 特征数据



步骤7 单击特征数据表下方的“数据探索”，用户可以在打开的数据分析文件内，通过“数据统计”页签查看所有特征的数据分布和数据分析详情信息，如图3-12所示。

图 3-12 数据统计详情



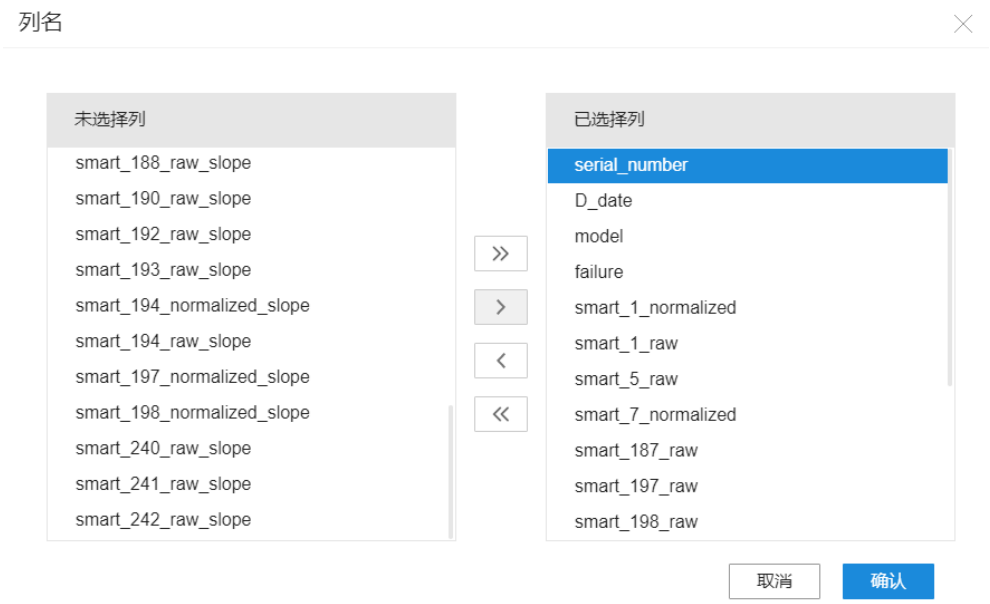
步骤8 单击特征工程操作主文件，返回特征工程编辑界面，在“算子”菜单的“数据处理”页签中选择“特征选择 > 选择列”。

步骤9 在“选择列”代码框中选择训练所需的特征列。

1. “列筛选方式”选择“列选择”。
2. 在“列名”框中单击“…”，在弹出的“列名”弹窗内选择如下特征列，如图3-13所示：

   serial_number，D_date，model，failure，smart_1_normalized，smart_1_raw，smart_5_raw，smart_7_normalized，smart_187_raw，smart_197_raw，smart_198_raw，smart_1_normalized_slope，smart_1_raw_slope，smart_5_raw_slope，smart_7_normalized_slope，smart_187_raw_slope，smart_197_raw_slope，smart_198_raw_slope

图 3-13 选择特征列



3. 单击“确认”。
   返回特征工程编辑界面，设置完成的“选择列”，如图3-14所示。

图 3-14 选择列

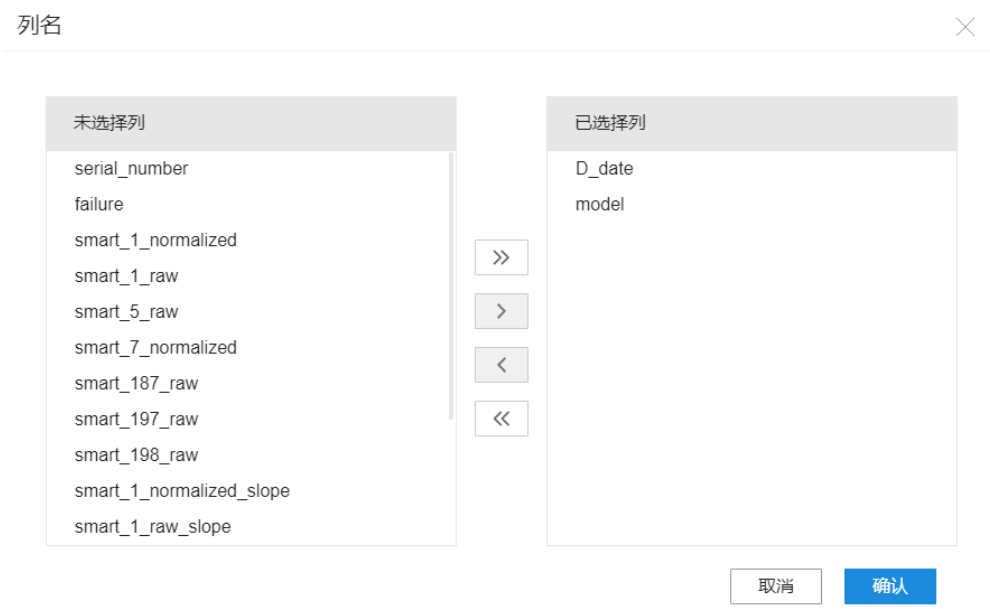
4. 单击“选择列”代码框的运行按钮，运行代码框。
   运行成功后“选择列”代码框下方，展示选定特征列的特征数据表。

步骤10 在“算子”菜单的“数据处理”页签中，选择“特征选择 > 删除列”。

步骤11 在“删除列”代码框中选择训练不需要的特征列。

1. “列筛选方式”选择“列选择”。
2. 在“列名”框中单击“...”，在弹出的“列名”弹窗内，选择如下特征列，如图3-15所示：
   D_date，model

图 3-15 选择待删除特征列



3. 单击“确认”。

   返回特征工程编辑界面，设置完成的“删除列”如图3-16所示。

图 3-16 删除列

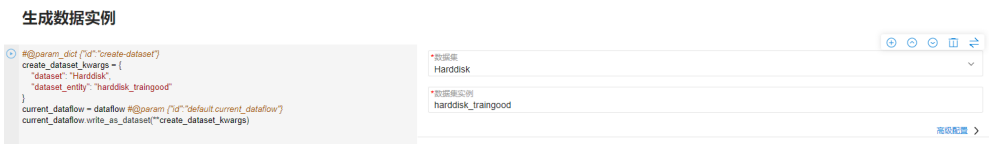
4. 单击“删除列”代码框的运行按钮，运行代码框。

   运行成功后“删除列”代码框下方，展示已删除选定特征列后的特征数据表。

步骤12 在“算子”菜单的“数据处理”页签中，选择“数据集 > 生成数据实例”，将特征操作流应用于绑定的数据，生成经过特征处理后的新数据。

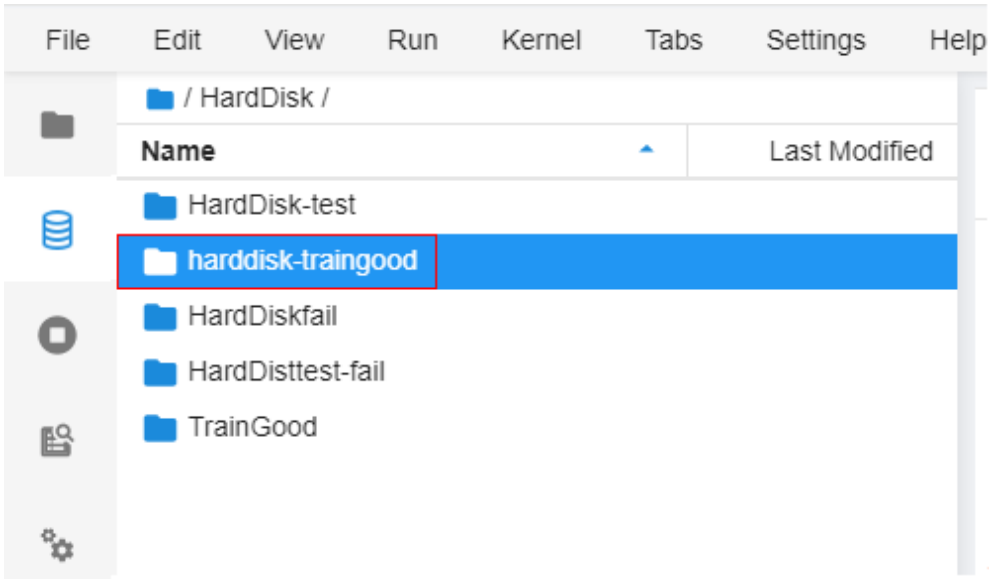
步骤13 在“生成数据实例”代码框中，选择数据集，并设置新数据实例名称，如图3-17所示。

图 3-17 生成全量数据实例



运行成功后，可在左侧目录中展开数据集目录，看到数据集目录下生成了新数据文件，如图3-18所示。

图 3-18 全量数据集

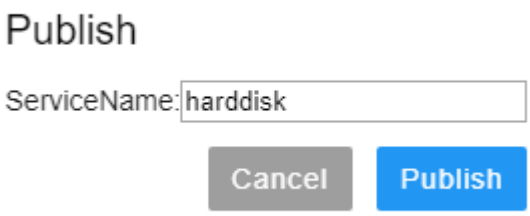


步骤14 单击特征工程菜单栏中的发布。

弹出如图3-19所示的对话框。

ServiceName：特征工程发布服务名，请根据实际情况设置。

图 3-19 特征工程服务



步骤15 单击“Publish”，将特征工程发布成服务。

步骤16 发布成功后，弹出“Success”弹窗，单击“OK”。

步骤17 单击“特征工程”，进入“特征工程管理”界面，单击“已发布服务”页签，查看特征工程服务，如图3-20所示。

图 3-20 已发布特征工程

步骤18 单击生成的特征工程服务所在行，对应“操作”列的  图标。

弹出如图3-21所示的对话框。

参数配置如下所示：

- 数据集：从下拉框中选择步骤1中创建的数据集。
- 数据实例：从下拉框中选择故障硬盘的训练数据集。
- 目标数据集：从下拉框中选择步骤1中创建的数据集。
- 目标数据实例：经过特征工程任务处理后，生成的数据集名称。请根据实际情况设置。
- AI引擎：AI算法运行平台，请选择“TF-1.8.0-python3.6”。
- 规格：资源配置信息，请按实际需求配置，如“2核|8G”。

图 3-21 创建任务

步骤19 单击“创建”，进入特征工程任务详情界面。

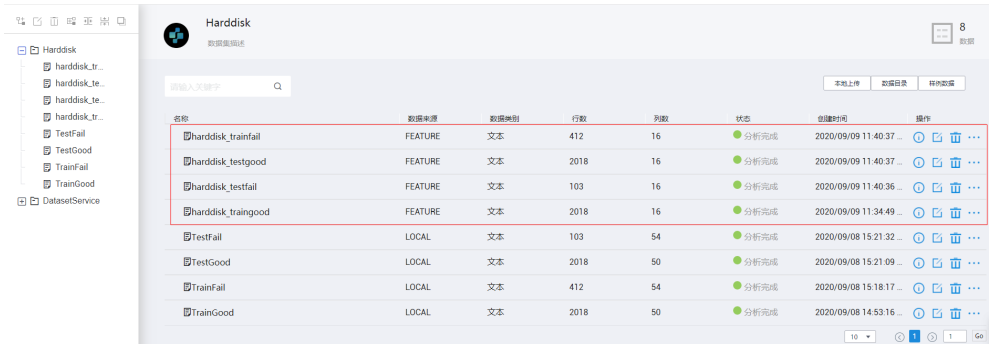
可以查看当前的任务进展。当任务的“任务状态”列为“FINISHED”时，说明故障硬盘训练集的特征处理操作完成。

步骤20 请参考步骤18~步骤19，依次对无故障硬盘测试和故障硬盘测试数据集做特征工程任务处理。

步骤21 单击菜单栏中的“数据集”。

在“数据集”界面，可以看到经过特征处理后，生成的四份数据，如图3-22所示。

图 3-22 数据详情



----结束

3.8 模型训练

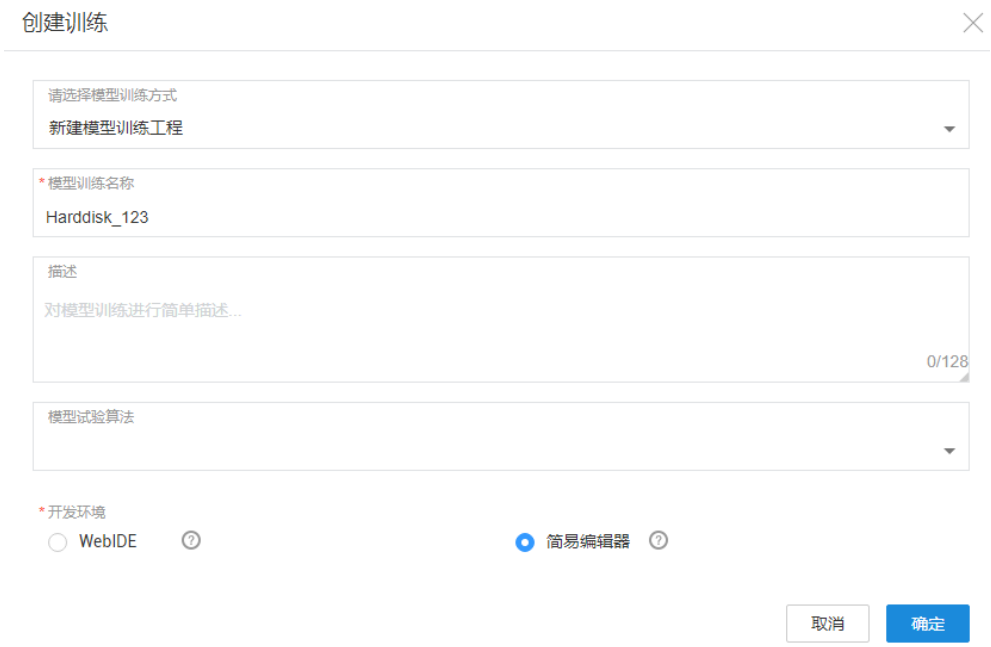
步骤1 单击菜单栏中的“模型训练”。

步骤2 单击"创建"，新建算法，如图3-23所示。

参数含义如下所示：

- 请选择模型训练方式：从下拉框中选择“新建模型训练工程”。
- 模型训练名称：按照界面提示进行设置。
- 开发环境：选择“简易编辑器”。

图 3-23 新建模型训练工程



步骤3 单击“确定”。

进入新创建的训练工程界面。

步骤4 单击界面右上角的图标。

进入训练代码编辑界面。

步骤5 单击展开代码目录，可在代码目录下根据实际需求添加代码文件，单击代码文件，在右侧编辑区域编辑代码。以下代码目录及文件创建仅以硬盘异常检测为例。

1. 单击项目根目录，单击“”，在根目录下创建代码文件夹“hardisk”。
2. 单击新建的文件夹“hardisk”，单击“”，在该文件夹下创建三个代码文件，分别为“__init__.py”，“preprocess.py”和“train.py”。
3. 将编辑好的代码分别拷贝进“preprocess.py”和“train.py”文件中，并按“Ctrl+S”保存。
4. 单击与训练工程同名的“.py”主入口文件，将编辑好的代码拷贝进主入口文件，并按“Ctrl+S”保存。

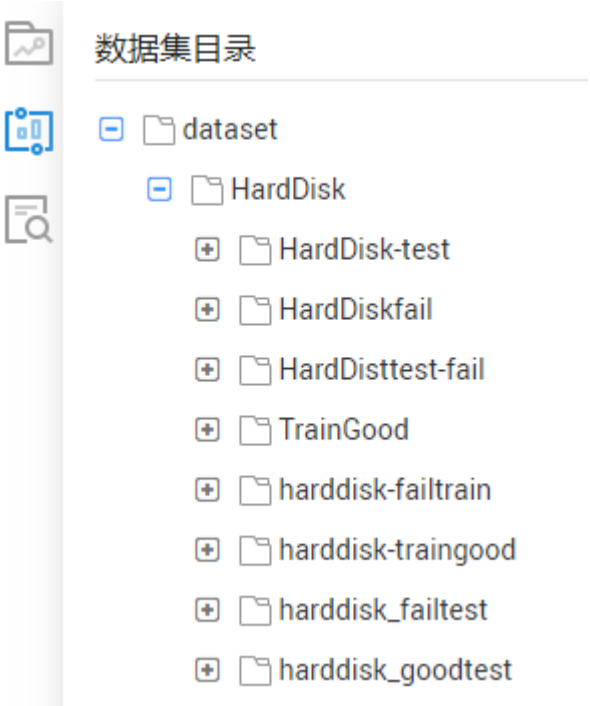
步骤6 单击代码目录左侧的，查看数据集目录，如图3-24所示。

Harddisk节点下面会展示原始导入的四份数据集和四份数据集分别经过特征处理后生成的数据集。

说明

“数据集目录”中展示的数据实例比“数据集”界面展示的数据实例多，属于正常，无需关注。

图 3-24 数据集

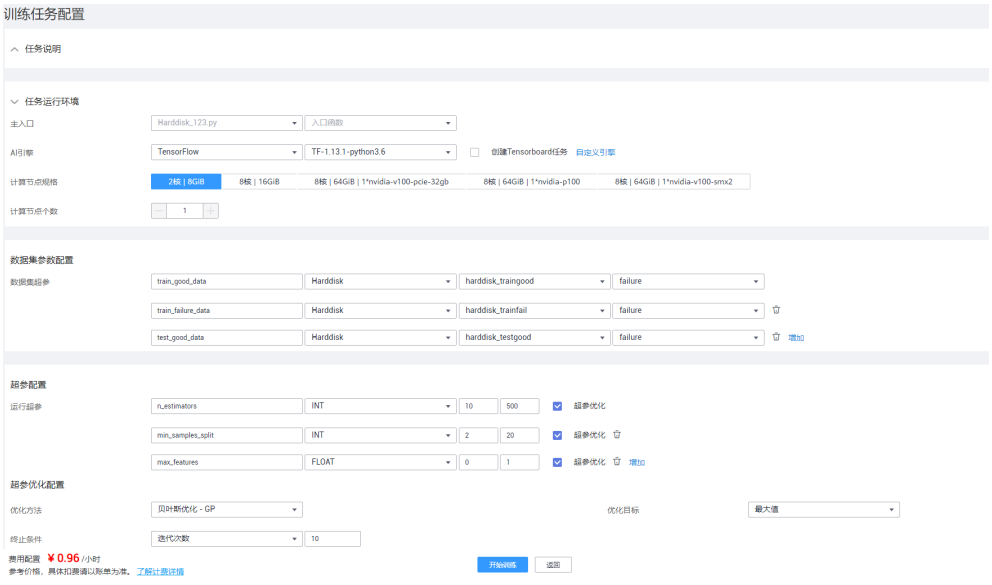


步骤7 单击“训练”，进入“训练任务配置”界面，配置训练任务，如图3-25所示。

参数含义如下所示：

- AI引擎：AI算法运行平台。从第一个下拉框中选择AI引擎“TensorFlow”，从第二个下拉框中选择匹配的python语言版本“TF-1.8.0-python3.6”。
- 计算节点规格：模型训练的资源配置信息。
- 计算节点个数：如果配置为“1”，表示使用1个节点进行训练；如果配置为2或者更大，表示使用分布式训练，开发者需要编写相应的调用代码。可使用内置的MoXing分布式训练加速框架进行训练，训练算法需要符合MoXing程序结构。可参考如下文档：https://github.com/huawei-clouds/modelarts-example/blob/master/moxing-apidoc/MoXing-API-UserInstructions.md
- 数据集超参：算法里面设置的所有数据集超参均展示在这里，一行对应一个超参。每行第一个方框中自动展示超参名称，需要从第二个和第三个下拉框中分别选择，超参对应的数据集和数据实例。第四个下拉框中中选择标签列“Failure”。其中，
  - train_good_data：设置为“数据集”界面创建的无故障硬盘训练数据集，经过特征处理后生成的数据集。
  - test_good_data：设置为“数据集”界面创建的无故障硬盘测试数据集，经过特征处理后生成的数据集。
  - train_failure_data：设置为“数据集”界面创建的故障硬盘训练数据集，经过特征处理后生成的数据集。
- 运行超参：模型参数是模型内部的配置变量，参数值可以根据数据自动估算。参数是机器学习的关键，通常从过去的训练数据中总结得出。超参区别于参数，是模型外部的配置，必须手工设置和调整，可用于帮助估算模型参数值。第一列是超参名称，第二列是超参的数据类型。如果勾选“超参优化”，则出现第三列和第四列，分别设置为当前超参取值范围的下限和上限。
- 优化方法：选择超参组合值的算法。保持默认值。
- 终止条件：超参优化停止的条件。保持默认值。如果选择“迭代次数”，则说明通过“贝叶斯优化”算法选取十个超参组合，分别进行模型训练。

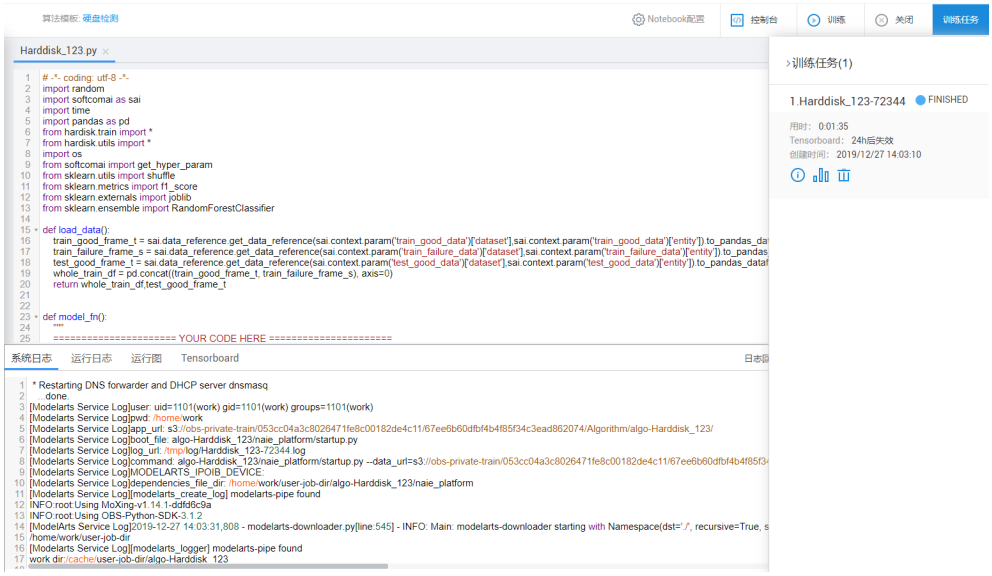
图 3-25 训练配置



步骤8 单击“开始训练”。训练任务的状态可通过单击"训练任务"查看。如图3-26所示。

加入训练后，下方自动展示模型训练日志、运行结果日志、运行图和Tensorboard窗口，也可以通过单击右上方“训练任务”，在展开的训练任务记录中单击 图标打开控制台窗口。

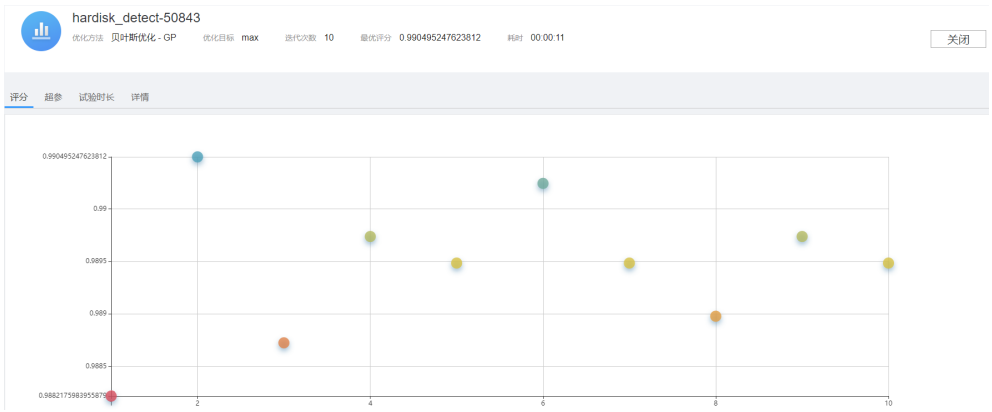
图 3-26 训练任务



模型训练结束后，单击 图标，可查看10个超参组合对应的10个模型训练评估结果，如图3-27所示。

- “评分”页签分别展示了10个模型训练任务的评分。
- “超参”页签分别展示了10个超参组合的取值。
- “试验时长”页签分别展示了10个超参组合对应的模型训练时长。
- “详情”页签分别展示了10个超参组合的迭代信息、耗时、评估值、超参取值，并支持对每个超参组合重新加入训练。

图 3-27 模型训练评估结果



步骤9 在评分页签内选取一个评分最高的模型任务数据，记录其三个超参值。参考**步骤7~步骤8**，配置最优模型的训练任务并进行训练。

该操作也可以用在**图3-27**的“详情”页签内，单击评分最高模型“操作列”对应的“  ”代替。

说明

对评分最高的模型再创建训练任务是为了在训练结束后，归档该最优模型包。模型训练任务在进行“超参配置”时，去勾选“超参优化”，三个超参值分别配置为此前记录的最优模型的三个对应超参值。

步骤10 单击菜单栏的“模型训练”。

进入模型训练界面。

步骤11 单击模型训练任务所在行。

进入模型训练任务详情界面。

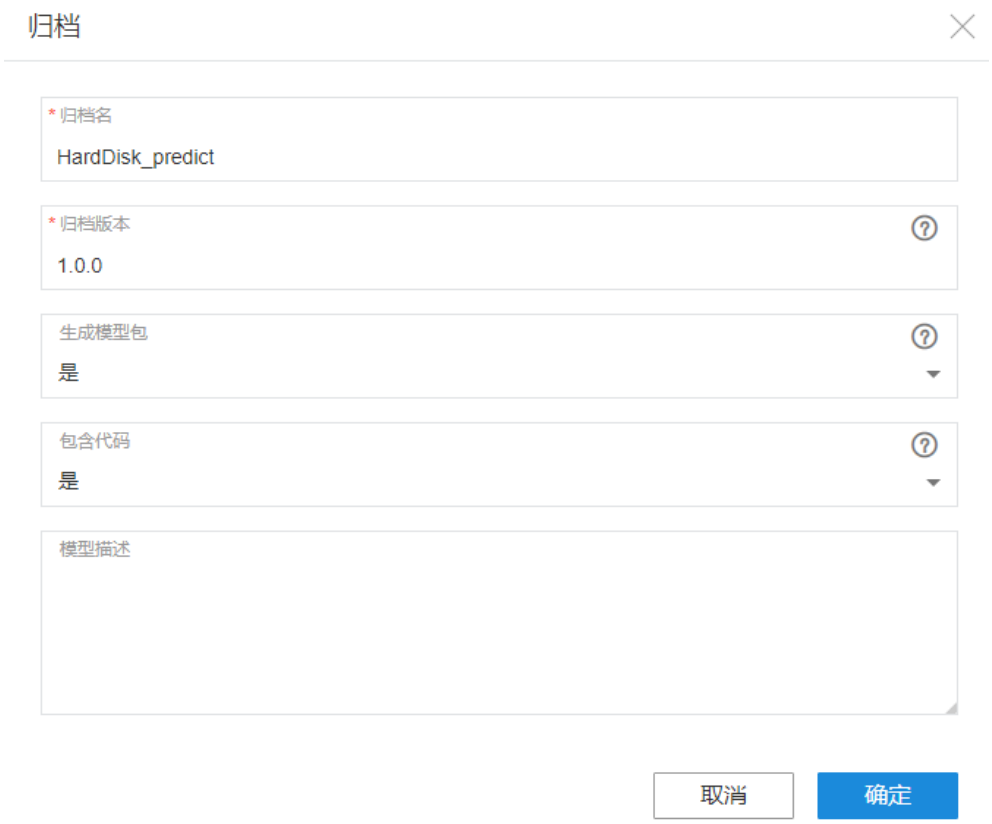
步骤12 在“模型训练任务”下面，单击最优模型生成的训练任务所在行的  图标。

弹出“归档”对话框，如**图3-28**所示。

主要参数说明如下：

- 生成模型包：是否直接在归档的同时打包模型包。选择“是”，表示同时对模型执行归档和打包操作；选择“否”表示仅对模型执行归档操作。默认选择“是”。
- 包含代码：模型包是否包含训练和推理的相关代码。选择“是”，表示包含，选择“否”，表示不包含。默认选择“是”。

图 3-28 模型归档



步骤13 单击“确定”。

----结束

3.9 模型管理

打包完成后，可在“模型管理”界面查看打包好的模型，如图3-29所示。

图 3-29 模型管理

3.10 模型验证

步骤1 单击菜单栏的“模型验证”。

步骤2 单击“创建”，弹出如图3-30所示的对话框。

其中，“模型类型”可以从下拉框中任选其一，无需勾选“创建模版验证代码”。

如果选择“Sklearn”，并勾选“创建模版验证代码”。则默认生成鸢尾花分类模型的验证代码。

图 3-30 创建验证服务

步骤3 单击“确定”。

进入当前验证服务所在界面。

步骤4 单击界面右上角的  图标。

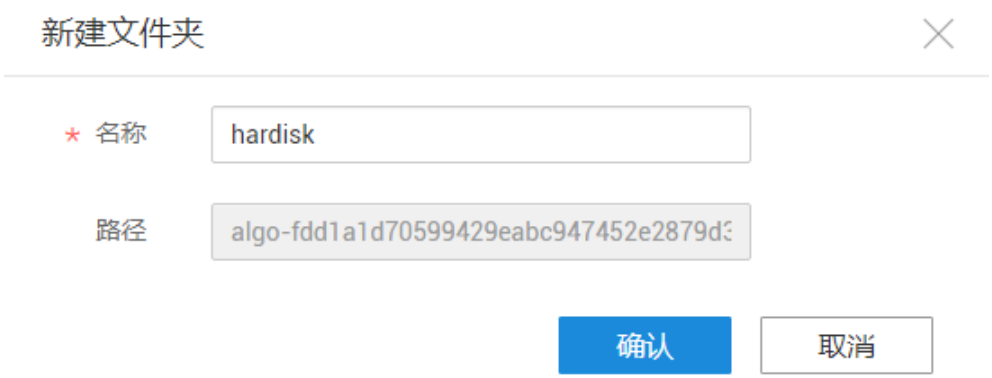
进入验证代码编辑界面。

步骤5 从本地拷贝已经编译好的代码至“validation.py”文件中，“Ctrl+S”保存文件。

步骤6 单击  ，新建文件夹。

文件夹名称命名为“hardisk”，如图3-31所示。

图 3-31 新建文件夹



步骤7 单击“确认”。

步骤8 选中“hardisk”，单击，新建算法文件。

文件名称命名为“utils.py”，如图3-32所示。

图 3-32 新建文件

步骤9 单击“确认”。

步骤10 在左侧导航栏中，单击“utils.py”打开文件，拷贝已经编译好的代码到此文件中，“Ctrl+S”保存文件。

步骤11 选中“hardisk”，单击，新建算法文件。

文件名称命名为“__init__.py”，该文件默认为空文件，用于标识python包。

步骤12 单击“确认”。

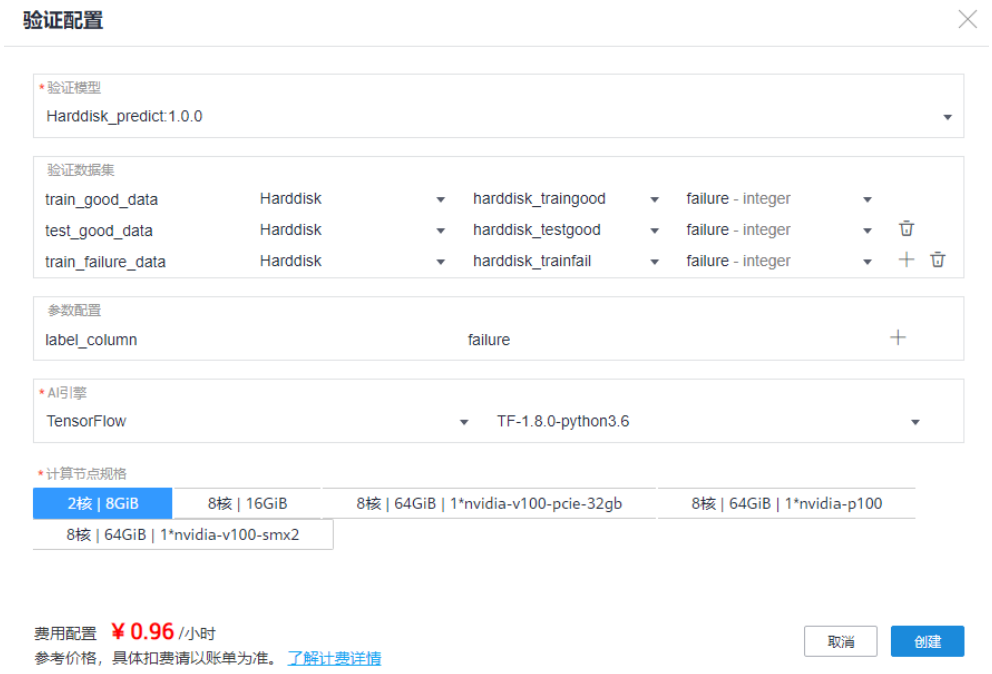
步骤13 单击“验证”，弹出“验证配置”对话框，如图3-33所示。

参数含义如下所示：

- 验证模型：从下拉框中选择“模型管理”中显示的模型。

- 验证数据集：一行对应一个验证数据集超参。每行第一个方框中输入超参名称，需要从第二个和第三个下拉框中分别选择，超参对应的数据集和数据实例名称。如果在“参数配置”中已经设置标签列，第四个下拉框可以置为空。其中，
  - train_good_data：设置为“数据集”界面创建的无故障硬盘训练数据集，经过特征处理后生成的数据集。
  - test_good_data：设置为“数据集”界面创建的无故障硬盘测试数据集，经过特征处理后生成的数据集。
  - train_failure_data：设置为“数据集”界面创建的故障硬盘训练数据集经过特征处理后生成的数据集。
- 参数配置：需要配置标签列参数“label_column”的值为“failure”。
- AI引擎：从第一个下拉框中选择AI引擎“TensorFlow”，从第二个下拉框中选择匹配的python语言版本“TF-1.8.0-python3.6”。
- 计算节点规格：模型训练的资源配置信息。

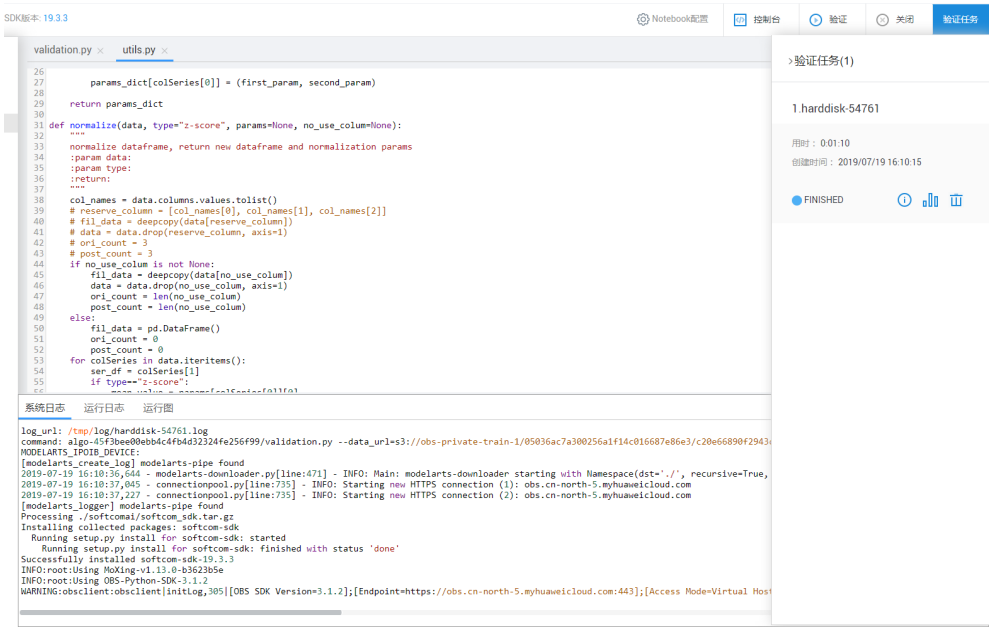
图 3-33 验证配置



步骤14 单击“创建”。验证任务可过单击“验证任务”查看，如图3-34所示。

加入验证后，下方自动展示模型验证日志、运行结果日志和运行图窗口，也可以通过单击“验证任务”，在展开的验证任务列表中单击 图标打开控制台窗口。

图 3-34 验证任务



模型验证结束后，单击  图标，可查看模型验证报告，如图3-35所示。

图 3-35 模型验证报告

----结束

3.11 云端推理

步骤1 单击菜单栏的“模型训练”，返回“模型训练”界面。

步骤2 单击训练工程对应的“  ”，进入模型训练工程编辑界面。

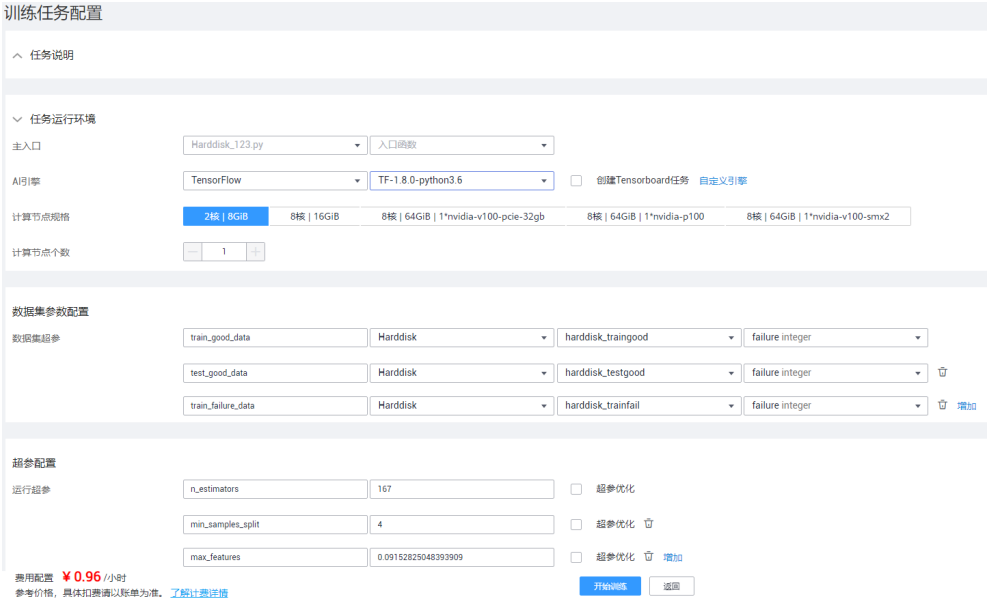
步骤3 在左侧代码目录区域单击“  ”，在工程根目录下创建训练代码文件“hardisk_detect_predict.py”。

步骤4 单击“hardisk_detect_predict.py”文件，将编辑好的推理代码拷贝入该文件，并按“Ctrl+S”保存。

步骤5 单击“  ”。

步骤6 设置训练任务参数，如图3-36所示。

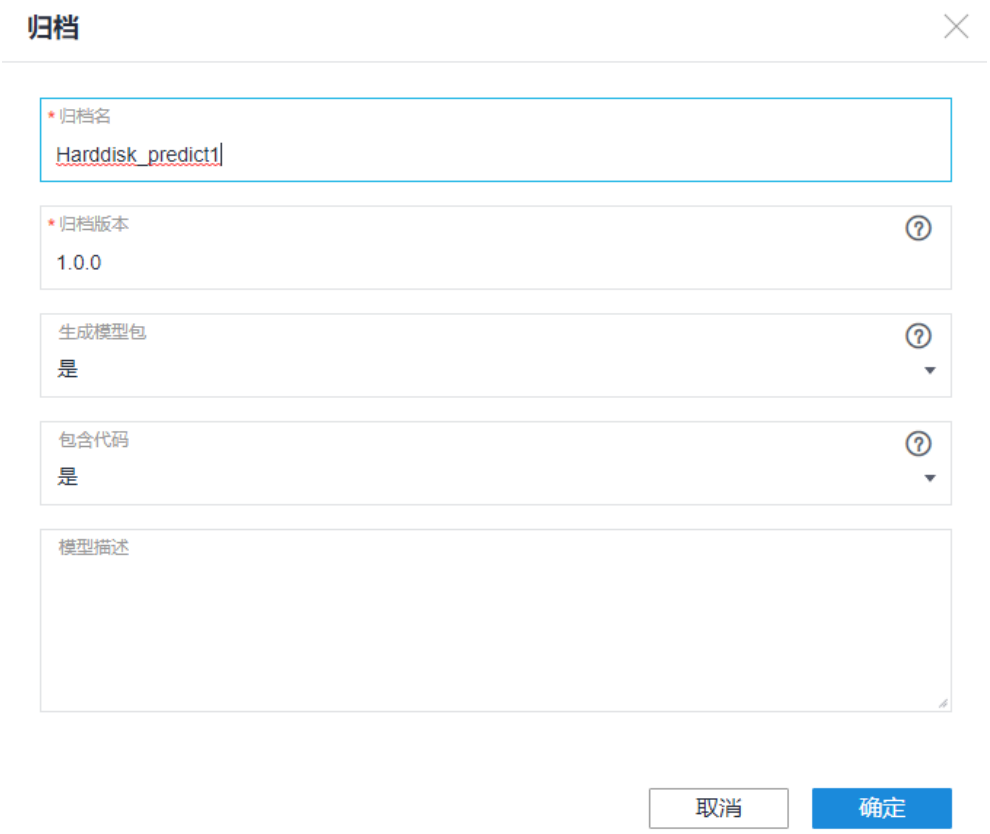
图 3-36 配置训练任务



步骤7 单击“开始训练”。

步骤8 待训练成功后，单击训练任务对应的“ ”，打包含推理代码的模型训练包，如图3-37所示。

图 3-37 打包推理模型包



步骤9 单击菜单栏的“模型管理”。

步骤10 单击“模型管理”界面右上角的“开发环境”，创建一个Webide环境，如图3-38所示。

图 3-38 创建 Webide 环境

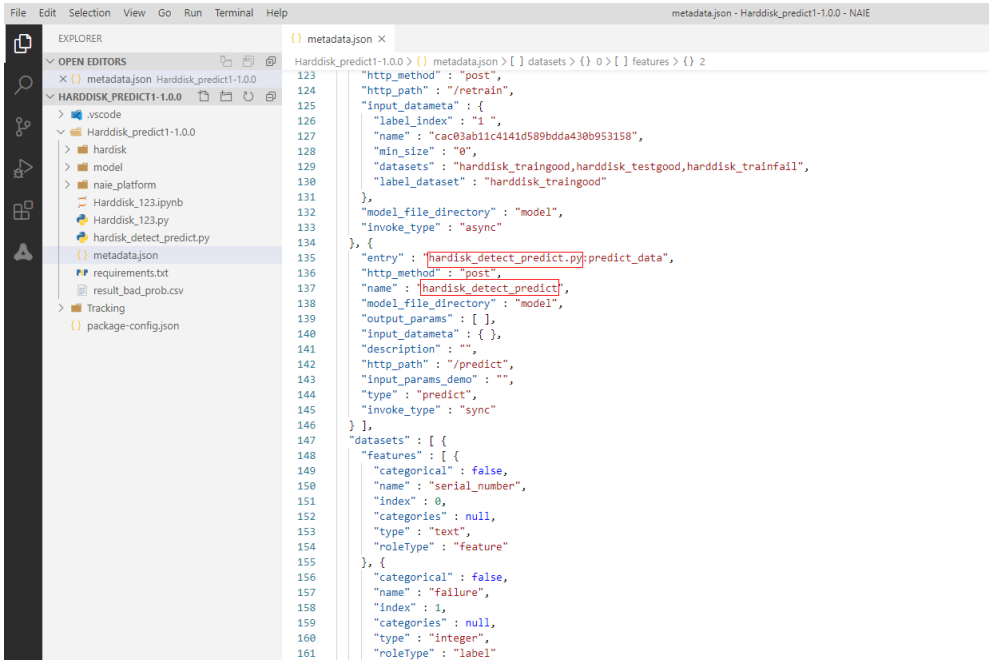
步骤11 待环境创建完毕后，单击步骤8模型包对应的“开发环境”，切换环境为新建的Webide环境，如图3-39所示。

图 3-39 切换 Webide 开发环境

步骤12 单击模型包操作列对应的“ ”，进入Webide代码编辑界面。

步骤13 单击左侧代码目录中的“metadata.json”文件，将红框内名字改成实际推理文件文字，如图3-40所示。

图 3-40 修改 metadata.json



步骤14 在Webide编辑界面左侧代码目录空白区域右键单击鼠标，选择“NAIE Package”。

步骤15 返回“模型管理”界面，单击模型包操作列对应的“  ”。

此处的模型包即为**步骤8**打包的模型包。

步骤16 在“发布推理服务”页面设置“版本”、“计算节点规格”等信息，单击“确定”，如**图3-41**所示。

等待系统发布推理服务，大约需要10分钟。

发布成功后，“  ”图标变为“  ”。

图 3-41 发布推理服务

步骤17 单击已发布推理服务模型包操作列对应的“ ”，进入推理服务的快速验证页面。

步骤18 在页面左侧“验证消息”区域中输入json格式的验证数据，单击“快速验证”，如图3-42所示。

验证数据样例如下：

```
{
  "smart_1_normalized": {
    "ZA19CLVQ": 0.176685,
    "ZA1A6RN7": -1.624761,
    "ZA1APLSW": -0.223636,
    "ZA1APWX6": 0.777167,
    "ZA1AQ5E2": -0.223636
  },
  "smart_1_raw": {
    "ZA19CLVQ": 0.218284,
    "ZA1A6RN7": -1.476697,
    "ZA1APLSW": -0.488849,
    "ZA1APWX6": 1.600456,
    "ZA1AQ5E2": -0.659933
  },
  "smart_5_raw": {
    "ZA19CLVQ": -0.12219,
    "ZA1A6RN7": -0.12219,
    "ZA1APLSW": -0.12219,
    "ZA1APWX6": -0.12219,
    "ZA1AQ5E2": -0.12219
  },
  "smart_7_normalized": {
    "ZA19CLVQ": -0.400716,
    "ZA1A6RN7": -1.372835,
```



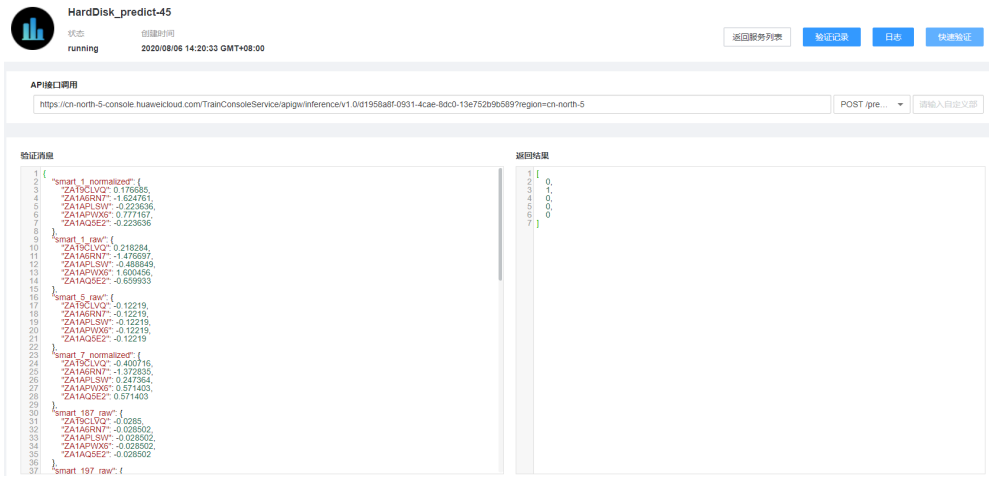
```
  "ZA1APLSW": 0.247364,
  "ZA1APWX6": 0.571403,
  "ZA1AQ5E2": 0.571403
},
"smart_187_raw": {
  "ZA19CLVQ": -0.0285,
  "ZA1A6RN7": -0.028502,
  "ZA1APLSW": -0.028502,
  "ZA1APWX6": -0.028502,
  "ZA1AQ5E2": -0.028502
},
"smart_197_raw": {
  "ZA19CLVQ": -0.113942,
  "ZA1A6RN7": -0.113942,
  "ZA1APLSW": -0.113942,
  "ZA1APWX6": -0.113942,
  "ZA1AQ5E2": -0.113942
},
"smart_198_raw": {
  "ZA19CLVQ": -0.113942,
  "ZA1A6RN7": -0.113942,
  "ZA1APLSW": -0.113942,
  "ZA1APWX6": -0.113942,
  "ZA1AQ5E2": -0.113942
},
"smart_1_normalized_slope": {
  "ZA19CLVQ": 1.235054,
  "ZA1A6RN7": -2.284543,
  "ZA1APLSW": 2.028689,
  "ZA1APWX6": 0.26889,
  "ZA1AQ5E2": 0.510431
},
"smart_1_raw_slope": {
  "ZA19CLVQ": 1.187602,
  "ZA1A6RN7": -3.581751,
  "ZA1APLSW": 0.022689,
  "ZA1APWX6": 0.506134,
  "ZA1AQ5E2": 0.060546
},
"smart_5_raw_slope": {
  "ZA19CLVQ": -0.107928,
  "ZA1A6RN7": -0.107928,
  "ZA1APLSW": -0.107928,
  "ZA1APWX6": -0.107928,
  "ZA1AQ5E2": -0.107928
},
"smart_7_normalized_slope": {
  "ZA19CLVQ": -0.254698,
  "ZA1A6RN7": 0.733461,
  "ZA1APLSW": 0.107928,
  "ZA1APWX6": 0.107928,
  "ZA1AQ5E2": 0.107928
},
"smart_187_raw_slope": {
  "ZA19CLVQ": -0.02716,
  "ZA1A6RN7": -0.02716,
  "ZA1APLSW": -0.02716,
  "ZA1APWX6": -0.02716,
  "ZA1AQ5E2": -0.02716
},
"smart_197_raw_slope": {
  "ZA19CLVQ": -0.063217,
```



```
    "ZA1A6RN7": -0.063217,
    "ZA1APLSW": -0.063217,
    "ZA1APWX6": -0.063217,
    "ZA1AQ5E2": -0.063217
  },
  "smart_198_raw_slope": {
    "ZA19CLVQ": -0.063217,
    "ZA1A6RN7": -0.063217,
    "ZA1APLSW": -0.063217,
    "ZA1APWX6": -0.063217,
    "ZA1AQ5E2": -0.063217
  }
}
```


右侧“返回结果”区域框会给出在线推理结果。

图 3-42 快速验证



----结束

3.12 修订记录

发布日期	修订记录
2020-09-30	框架切换，全篇更换截图。 优化“云端推理”章节。
2020-08-17	新增“云端推理”章节。 修改“模型管理”、“模型验证”章节截图。
2020-07-16	Jupyterlab优化，对应特征工程章节截图更新。 模型训练界面优化，对应模型训练章节截图更新。

发布日期	修订记录
2020-06-30	模型管理界面新增推理服务入口、新增创建联邦学习案例入口，对应模型管理章节截图更新。 Jupyterlab算子菜单位置及算子分组变更，对应特征工程章节菜单入口描述变更。 Jupyterlab特征工程选择数据增加时序数据选择，并支持多数据选择，对应特征工程章节操作截图全量更新。
2020-03-30	训练平台界面优化，训练服务操作界面截图全量更新。
2019-12-30	快速入门从鸢尾花分类建模变更为硬盘异常检测建模，资料全部重新写作。
2019-04-30	第一次正式发布。

4 用户指南

4.1 文档导读

本文档包含了使用模型训练服务前的准备工作、如何使用训练平台导入数据、特征操作、模型训练、模型打包、模型验证以及云端推理框架的操作指导，用户可以根据**文档导读**查找需要的内容。

表 4-1 文档导读

阶段	章节
了解模型训练服务	**训练服务简介**
训练平台的操作流程简介及访问服务的流程	• **操作流程** • **访问训练平台**
熟悉训练平台中数据集、特征工程、模型训练、及模型管理相关操作	• **项目创建** • **数据集** • **特征工程** • **模型训练** • **模型管理**
在线对训练模型进行测试验证	**模型验证**
模型发布成服务后，在线验证模型推理效果	**云端推理框架**

4.2 训练服务简介

模型训练服务为开发者提供电信领域一站式模型开发服务，从数据预处理，到特征提取、模型训练、模型验证、在线推理，本服务为开发者提供开发环境、模拟验证环境，API和一系列开发工具，帮助开发者快速高效开发电信领域模型。

电信经验嵌入降低模型开发门槛

- 集成50+电信领域AI算子&项目模板提升训练效率，降低AI开发门槛，让开发者快速完成模型开发和训练
- AutoML自动完成特征选择、超参选择及算法选择，提升模型开发效率
- 高效开发工具JupyterLab和WebIDE：交互式编码体验、0编码数据探索及云端编码及调试

联邦学习&重训练，保障模型应用效果

- 支持联邦学习，模型可以采用多地数据进行联合训练，提升样本多样性，提升模型效果
- 支持迁移学习，只需少量数据即可完成非首站点模型训练，提升模型泛化能力
- 模型自动重训练，持续优化模型效果，解决老化劣化问题

预置多种高价值通信增值服务，缩短模型交付周期

- 无需AI技能，支持模型自动生成，业务人员快速使用
- 多种通信增值服务开箱即用，快速支撑电信领域AI应用

支持 3 种部署模式

- 公有云部署：数据允许出局，面向用户包括：中小T、合作伙伴、华为内部研发。
- 合营云部署：数据不出局，面向用户为有合营云的大T。
- 华为云Stack部署：数据不出局，面向用户为无合营云的大T。

4.3 准备工作

4.3.1 订购模型训练服务

步骤1 在用户终端PC上打开浏览器，在地址栏中输入“**https://console.huaweicloud.com/naie/**”，进入AI市场。

用户首次访问AI市场，会进入“访问授权”界面，单击“授权并继续”即可。

步骤2 单击界面右上角的“登录”，进入登录界面。

步骤3 输入租户名和密码，单击“登录”，进入AI市场。

首次登录后请及时修改密码，并定期修改密码。

步骤4 依次选择“AI服务 > 模型与训练服务 > 模型训练 > 模型训练服务”，进入模型训练服务介绍页面。

步骤5 单击“我要购买”，进入如**图4-1**所示的界面。

区域：为用户提供服务的华为云Region。请选择“华北-北京四”。

用户可以单击“了解计费详情”，详细了解训练服务提供的资源、规格和相应的价格信息。同时，用户在使用具体资源时，训练服务会在界面给出醒目的计费提示。

图 4-1 订购训练服务

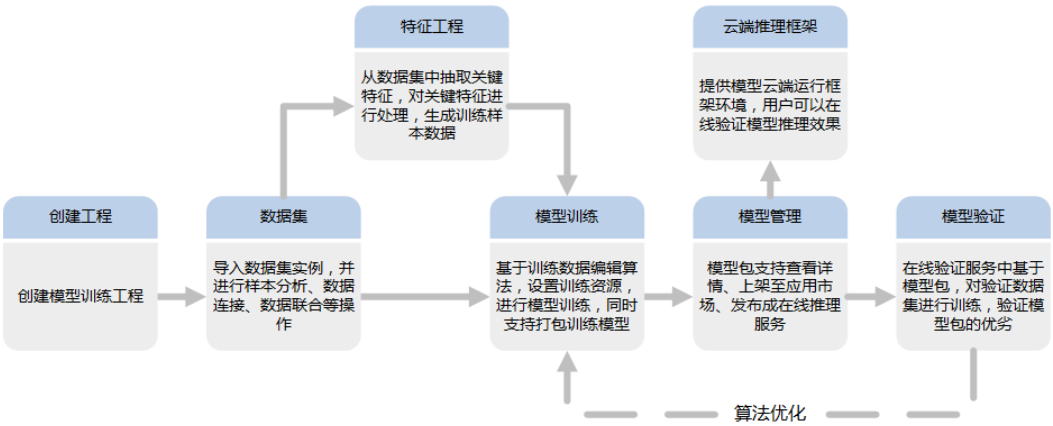
步骤6 单击“立即使用”，服务订购完成。

----结束

4.3.2 操作流程

训练服务为用户提供了数据集、特征处理、模型训练、模型管理、模型验证以及云端推理框架能力，用户使用流程如图4-2所示。

图 4-2 训练服务操作流程



4.3.3 访问模型训练服务

步骤1 在用户终端PC上打开浏览器，在地址栏中输入“https://console.huaweicloud.com/naie/”，进入AI市场。

步骤2 单击界面右上角的“登录”，进入登录界面。

步骤3 选择“IAM用户登录”方式，输入租户名、用户名和密码。

用户也可以直接通过账号登录。首次登录后请及时修改密码，并定期修改密码。

步骤4 单击“登录”，进入AI市场。

步骤5 依次选择“AI服务 > 模型与训练服务 > 模型训练 > 模型训练服务”，进入模型训练服务介绍页面。

步骤6 单击“进入服务”，进入模型训练服务页面。

----结束

4.4 项目创建

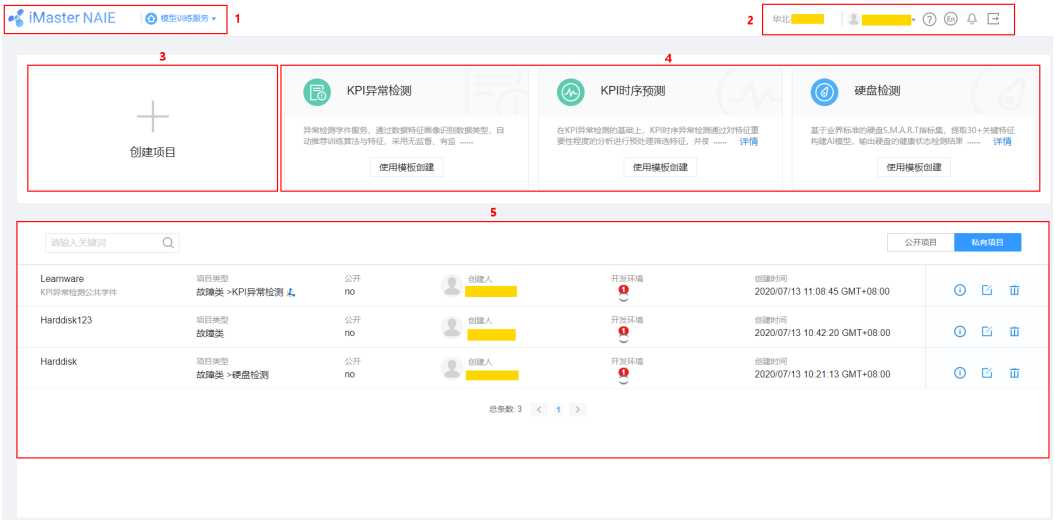
4.4.1 训练服务首页简介

训练服务首页展示了用户自己创建的项目和用户所属租户下面其他用户创建的公开项目，提供如下功能：

- 创建项目
- 使用模板快速创建项目，模板中已经预制数据集、特征处理算法、模型训练算法和模型验证算法。
- 查看和编辑项目信息

训练服务首页界面如下图所示。

图 4-3 训练服务首页



训练服务首页介绍如表4-2所示。

表 4-2 训练服务首页说明

区域	参数名称	参数说明
1	iMaster NAIE	当前服务所属的品牌名称。
	模型训练服务	单击服务名称图标下拉框，从下拉框中选择服务名称，可以进入对应服务的首页界面。
2	华北-北京一	用户账户所属Region。

区域	参数名称	参数说明
		当前用户的头像和用户名。 单击用户名右侧的倒三角图标，可查看当前用户创建的所有开发环境和TensorBoard环境，功能说明如下所示： • 开发环境：支持启动、停止或删除开发环境（Jupyterlab、WebIDE和Notebook环境）。 • TensorBoard：单击“TensorBoard”，可查看TensorBoard环境列表。单击环境列表中的TensorBoard环境名称，可跳转到相应的训练任务。
		帮助中心快捷入口。
		训练服务中英文界面切换按钮。
		用户创建项目的通知信息，包括数据集、特征工程、模型训练、模型管理和模型验证中任务执行失败的所有通知。
		退出登录图标。
3		创建项目图标。
4	• KPI异常检测 • KPI时序预测 • 硬盘检测	训练服务预置的网络领域开发模板，可以直接单击“使用模板创建”，生成对应领域的项目，项目中预置了数据集、特征工程操作流、模型训练算法和模型验证算法。
5		搜索项目名称关键字，快速查找项目。
		用户创建项目的时候，选择公开给指定的用户组，则用户组内的所有用户均可见和使用。
		用户创建项目的时候，选择不公开，则仅当前用户可见和使用。
	Walkthroughs_55068	项目名称。
	项目类型	项目分类。 包含如下选项： • 故障类 • 能源利用 • 资源利用 • 用户体验 • 其他

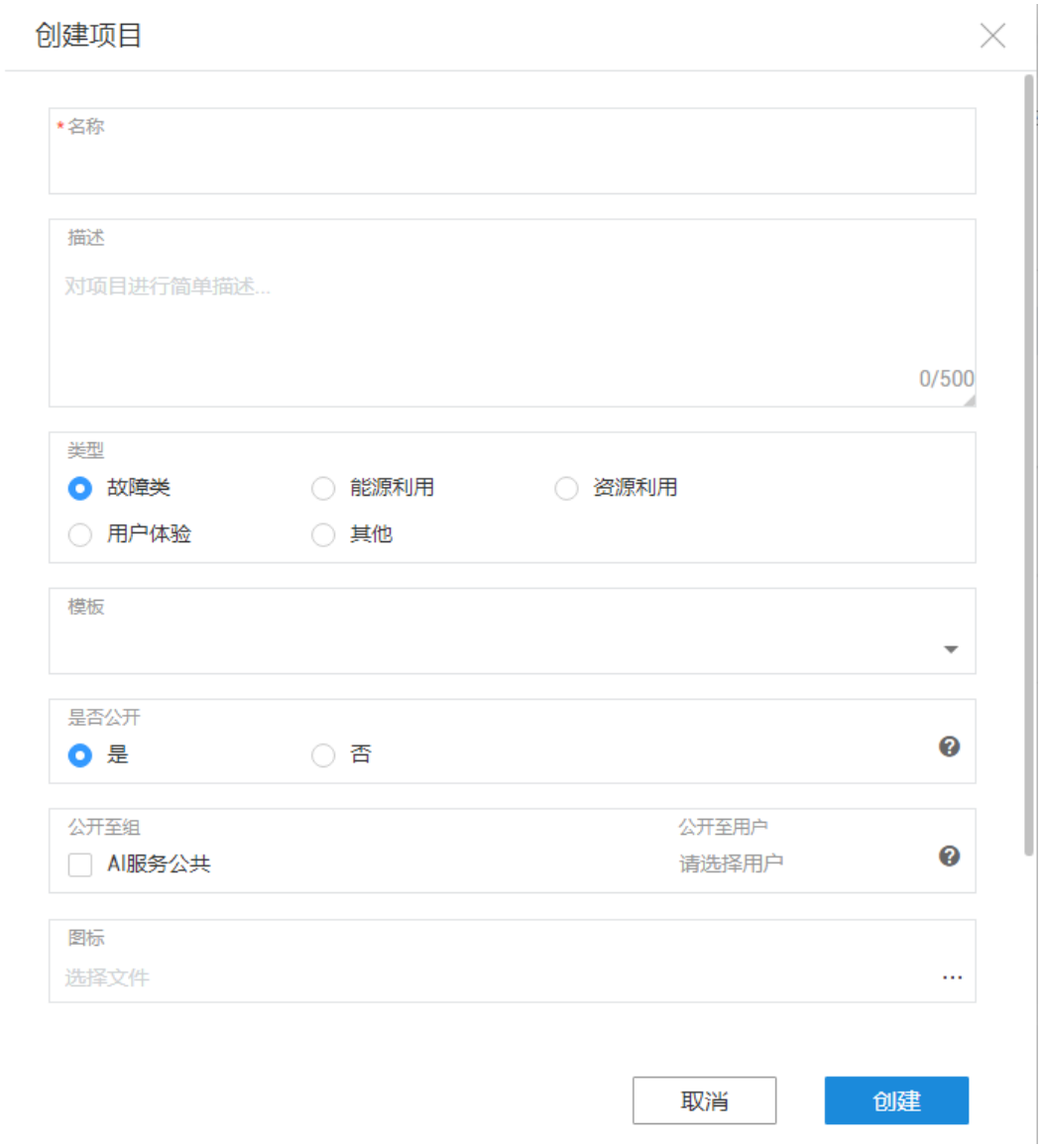
区域	参数名称	参数说明
	公开	项目是否公开给当前租户下的其他用户查看和使用。 包含如下选项： • yes • no
	创建人	创建项目的用户头像和用户名。
	开发环境	分类展示当前项目创建的Jupyterlab、WebIDE和普通的Notebook环境数量。 单击Jupyterlab、WebIDE或Notebook图标，打开当前项目下对应类型的开发环境信息，弹窗中单击“更多”，可以查看其它类型的开发环境列表。
	创建时间	项目创建时间。
		进入项目总览页面。
		支持修改如下项目的信息： • 描述 • 是否公开 • 自定义项目图标 如果置灰，表示您不是当前项目的创建者，没有权限修改项目信息。
		删除项目。 如果置灰，表示您不是当前项目的创建者，没有权限删除项目。

4.4.2 创建项目

使用训练服务进行模型训练前，需要先创建一个项目。训练平台会提供一定的计算资源给每个项目。

步骤1 在训练平台首页，单击“创建项目”上方的“+”按钮，弹出“创建项目”对话框，如图4-4所示。

图 4-4 创建项目



步骤2 配置“创建项目”对话框参数，如表4-3所示。

表 4-3 参数说明

参数名称	参数说明
名称	项目的名称。 名称只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线（_）（-）组成，不能以下划线和中划线结尾，且长度为[2-20]个字符。
描述	对项目的简要描述。 字数不能超过500。

参数名称	参数说明
类型	创建项目的类型，包括以下几种： ● 故障类 ● 能源利用 ● 资源利用 ● 用户体验 ● 其他
模板	已有网络领域经验的沉淀，复用已有的网络经验项目。使用模板创建项目后，项目中会预置有相关的数据集、特征处理操作流、模型训练算法以及模型验证算法。当前支持的模板有： ● KPI异常检测 ● KPI时序检测 ● 硬盘检测
是否公开	项目是否可以被所属用户组的其他用户访问： ● 是 ● 否
公开至组	仅当“是否公开”设置为“是”，才会展示“公开至组”。 默认展示当前用户所属的所有用户组，如果勾选用户所属的用户组，则被勾选用户组下的所有用户均可以查看当前项目。
图标	项目图标。 支持用户本地上传。

步骤3 单击“创建”，完成训练平台项目的创建。

----结束

4.4.3 项目概览

项目概览界面显示了当前项目的信息概览，如图4-5所示。

图 4-5 项目概览界面

项目概览界面包括：

- 数据处理、训练状态统计、训练服务、验证服务的运行状态。
- 数据集、特征工程、模型训练、模型管理、模型验证的列表信息。用户可以直接单击对应列表中的“创建”，创建新的功能模块。
- 项目最新操作的通知信息。

4.5 数据集

4.5.1 数据集简介

基本概念

数据集模块主要为训练平台提供统一的数据管理能力，数据集可以提供给特征工程，做特征处理和提取关键特征；也可以直接加入模型训练。数据集相关的两个基本概念：

- 数据集：某业务下具有相同数据格式数据的逻辑集合。
- 数据：数据实例，有具体的特征和样本数据。

数据集以文件夹的形式管理数据，一个数据集中可以包含多份数据，从而对数据进行高效简洁的管理。用户可以根据数据的业务特点建立数据集，例如在大型DC PUE Case中，可以创建空调、冷站等数据集，再分别创建相应的数据。

数据来源

数据实例来源有五种：

- 从用户本地上传
- 导入样例数据
- 从数据目录订阅

- 数据经过特征处理并应用特征操作流后，系统自动生成的数据
- 数据连接或数据合并后，系统自动生成的目标数据

操作说明

导入数据的方式包括本地上传、导入样例数据、订阅数据目录数据三种。经特征处理、数据连接或数据联合后生成的数据，为系统自动生成的数据，不支持用户手动导入。

“数据集”在创建数据集、导入数据后，还支持对数据进行分析。用户可以根据数据结果对数据质量进行评估，判断数据集是否可以直接进行模型训练，或必须经过特征处理后才能加入模型训练。“数据集”还支持将多份数据进行“数据联合”或“数据连接”的操作，用于增强样本或扩展特征维度。数据集相关操作请参见**数据集操作**。

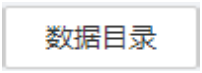
数据集页面

“数据集”页面包含了“数据目录”和“数据集详情”两个区域框。在“数据目录”区域框中，可以新建数据集、导入数据集的数据实例、删除数据。在“数据详情”区域框，可以通过列表的形式查看数据详情、对数据执行特征工程、基于数据新建特征工程、跳转模型训练界面、删除数据。“数据集”页面详情请参见**图4-6**，“数据集”页面操作信息，请参见**表4-4**。

图 4-6 数据集页面

表 4-4 数据集列表说明

区域	参数名称	参数说明
1		新建数据集。
		修改数据集别名和描述。需要先选中数据集，再修改数据集信息。
		删除数据集或数据。
		导入数据。 当前支持本地上传、样例数据、数据目录三种方式。

区域	参数名称	参数说明
		数据连接。详情请参见**数据连接**。
		数据联合。详情请参见**数据联合**。
		数据同步图标。用户在数据集服务上订阅完成数据集后，支持一键式导入至训练服务的DatasetService数据集中。
2		输出数据名称关键字，快速检索数据。
		本地上传数据的快捷入口。
		通过数据目录订阅数据的快捷入口。
		选择样例数据的快捷入口。
3	名称	数据实例的名称。
	数据来源	数据实例的来源： ● LOCAL：用户本地上传的数据。 ● SAMPLE：样例数据。 ● DATACATALOG ： 在数据目录中订阅的数据。 ● FEATURE：数据经过特征处理并应用特征操作流后，自动生成的数据。单击“FEATURE”，可跳转至对应的特征处理界面。 ● MERGE：数据连接或数据合并后，系统自动生成的目标数据。
	数据类别	导入数据的类别。 包含如下选项： ● 文本 ● 图片 ● 其他
	行数	数据的样本数量。
	列数	数据的特征列数量。
	状态	数据的当前状态。
	创建时间	数据创建的时间。

区域	参数名称	参数说明
	操作	可对数据执行的操作： •  ：查看数据详情。 •  ：修改数据信息，包括：实例别名、数据类型、文件编码、分隔符、标题行。 •  ：删除数据。 •  ：对数据执行已有特征工程的操作流，并生成新的数据。特征工程操作请参见特征工程。特征工程处理过的数据，不能再用相同的特征工程进行二次处理。 •  ：使用当前数据创建新的特征工程。创建特征工程的方法请参见创建特征工程。 •  ：跳转至“模型训练”页面。模型训练操作请参见模型训练。

DatasetService 数据集

DatasetService数据集是模型训练服务预置的数据集，专门存放从数据集服务订阅的数据集。

操作方法为：用户在数据集服务订阅完成数据集，回到模型训练服务的数据集菜单界面，单击左侧的DatasetService数据集，再单击右侧界面右上角的“数据同步”，实现数据集服务订阅的数据一键式导入训练服务，如图4-7所示。

图 4-7 DatasetService 数据集

4.5.2 新建数据集和导入数据

用户根据数据的业务类别创建数据集，并导入数据。

基本功能介绍

系统支持上传本地数据、从公共空间中导入样例数据、从数据目录中订阅数据集。操作步骤如下。

步骤1 单击“项目总览”页面“数据集”下方的“创建”。

进入“数据集”页面，弹出“导入数据”对话框，如图4-8所示。

图 4-8 导入数据

配置“导入数据”对话框参数，具体参见表4-5。

表 4-5 导入数据参数说明

参数名称	参数说明
数据集	支持用户编辑生成新的数据集，示例：Harddisk。

参数名称	参数说明
数据类别	导入数据的类别。 包含如下选项： ● 文本 ● 图片 ● 其他
实例名称	数据实例的名称。 只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线（_）、（-）组成，不能以下划线或中划线结尾，且长度为[1-128]个字符。
实例别名	数据实例的别名。 由字母、汉字、数字（0~9）、下划线（_）、中划线（-），圆括号组成，且长度为[1-128]个字符。创建别名后，系统将优先以数据集别名显示数据集。
数据来源	数据上传的途径。 包含如下方式： ● 本地上传：从用户本地上传数据。 ● 数据目录：导入用户在数据服务的数据集服务中订阅的数据。 ● 样例数据：训练平台环境中预置的用户体验数据。包括鸢尾花原始测试集、鸢尾花训练集、鸢尾花测试集、KPI 15分钟数据集、KPI 60分钟数据集、KPI异常检测数据集。 其中鸢尾花原始测试集、KPI 15分钟数据集和KPI 60分钟数据集中包括空值，用户可以通过特征工程进行数据修复，剔除空值。
本地上传-文件大小限制为80M，文本支持csv和txt	数据来源选择“本地上传”时可见，表示数据文件所在的用户本地路径。 为避免后续处理数据时出错，请按要求上传csv和txt格式的数据文件。

参数名称	参数说明
数据目录-请选择数据集	数据来源选择“数据目录”时可见。 选择数据集服务中订阅的数据。 ● 订阅：单击“订阅”图标，自动跳转至数据湖的数据集服务界面，可以查询并订阅数据。 ● 刷新：刷新展示数据集服务订阅的数据列表。 ● 数据名称：数据集服务订阅的数据名称。 ● 申请状态：数据集服务订阅数据的申请状态。 ● 审批人：数据集服务订阅数据的审批责任人。 ● 数据来源：数据集服务订阅数据的来源。 **说明** 在订阅数据目录数据前，需要用户阅读《使用须知》，并签署同意遵守使用敏感数据项目条款或条件约束。
样例数据-请选择数据集	数据来源选择“样例数据”时可见。 系统默认给出六个数据实例： ● iris_raw：鸢尾花原始测试集 ● iris_training：鸢尾花训练集 ● iris_test：鸢尾花测试集 ● KPI_15mins：KPI 15分钟数据集 ● KPI_60mins：KPI 60分钟数据集 ● TPC-iSPS11_60：KPI异常检测数据集 其中，iris_raw、KPI_15mins、KPI_60mins数据集中包含空值。用户可以通过特征工程进行数据修复，剔除空值。
分隔符	用户根据导入数据文件的格式进行选择，用于系统识别数据字段。 当前支持“,”、“;”和“|”三种分隔符。
文件编码	数据文件的编码格式。 当前支持UTF-8、GBK和GB2312三种格式。
标题行	数据是否包含标题行，用户根据导入数据文件的格式进行选择。 包含如下选项： ● 有标题行 ● 无标题行

步骤2 单击“创建”，导入数据文件。

如果导入数据所在的“状态”列显示“导入成功”，说明数据导入成功。

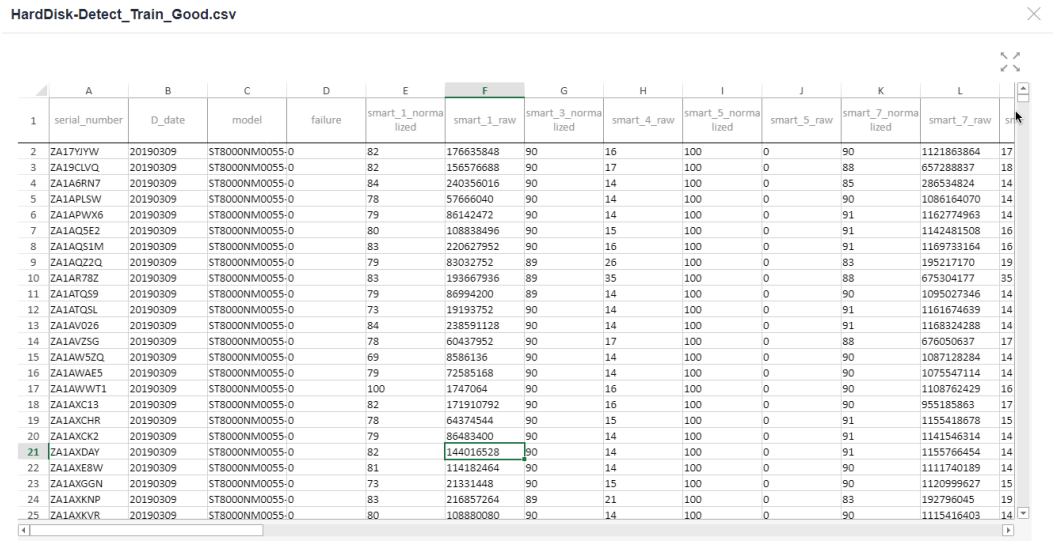
步骤3 单击数据实例所在行对应“操作”列的图标，进入数据详情界面，如图4-9所示。

图 4-9 数据详情

步骤4 单击数据集所在行对应“操作”列的“查看”，可以查看数据内容，如所图4-10示。

单击数据集所在行对应“操作”列的“删除”，可以删除当前数据集。

图 4-10 数据内容



	A	B	C	D	E	F	G	H	I	J	K	L
1	serial_number	D_date	model	failure	smart_1_normalized	smart_1_raw	smart_3_normalized	smart_4_raw	smart_5_normalized	smart_5_raw	smart_7_normalized	smart_7_raw
2	ZA17YJYW	20190309	ST8000NM0055-0		82	176635848	90	16	100	0	90	1121863864
3	ZA19CLVQ	20190309	ST8000NM0055-0		82	156576688	90	17	100	0	88	657288837
4	ZA1A6RN7	20190309	ST8000NM0055-0		84	240356016	90	14	100	0	85	286534824
5	ZA1APLSW	20190309	ST8000NM0055-0		78	57666040	90	14	100	0	90	1086164070
6	ZA1APWX6	20190309	ST8000NM0055-0		79	86142472	90	14	100	0	91	1162774963
7	ZA1AQ5E2	20190309	ST8000NM0055-0		80	108838496	90	15	100	0	91	1142481508
8	ZA1AQS1M	20190309	ST8000NM0055-0		83	220627952	90	16	100	0	91	1169733164
9	ZA1AQZ2Q	20190309	ST8000NM0055-0		79	83032752	89	26	100	0	83	195217170
10	ZA1AR78Z	20190309	ST8000NM0055-0		83	193667936	89	35	100	0	88	675304177
11	ZA1ATQS9	20190309	ST8000NM0055-0		79	86994200	89	14	100	0	90	1095027346
12	ZA1ATQSL	20190309	ST8000NM0055-0		73	19193752	90	14	100	0	91	1161674639
13	ZA1AV026	20190309	ST8000NM0055-0		84	238591128	90	14	100	0	91	1168324288
14	ZA1AVZSG	20190309	ST8000NM0055-0		78	60437952	90	17	100	0	88	676050637
15	ZA1AW5ZQ	20190309	ST8000NM0055-0		69	8586136	90	14	100	0	90	1087128284
16	ZA1AWAE5	20190309	ST8000NM0055-0		79	72585168	90	14	100	0	90	1075547114
17	ZA1AWWT1	20190309	ST8000NM0055-0		100	1747064	90	16	100	0	90	1108762429
18	ZA1AXC13	20190309	ST8000NM0055-0		82	171910792	90	16	100	0	90	955185863
19	ZA1AXCHR	20190309	ST8000NM0055-0		78	64374544	90	15	100	0	91	1155418678
20	ZA1AXCK2	20190309	ST8000NM0055-0		79	86483400	90	14	100	0	91	1141546314
21	ZA1AXDAY	20190309	ST8000NM0055-0		82	144016528	90	14	100	0	91	1155766454
22	ZA1AXE8W	20190309	ST8000NM0055-0		81	114182464	90	14	100	0	90	1111740189
23	ZA1AXGGN	20190309	ST8000NM0055-0		73	21331448	90	15	100	0	90	1120999627
24	ZA1AXKNP	20190309	ST8000NM0055-0		83	216857264	89	21	100	0	83	192796045
25	ZA1AXKVR	20190309	ST8000NM0055-0		80	108880080	90	14	100	0	90	1115416403

步骤5 单击图4-9中的“元数据”，进入数据分析界面，如图4-11所示。

图 4-11 数据分析

步骤6 请根据实际情况，从下拉框中选择AI引擎和对应的规格，单击“数据分析”。

可查看数据实例的详细信息，包括字段名称、字段类型、数据分布、有效值、空值、异常值、最大值、最小值、均值、方差、分位数等，如图4-12所示。

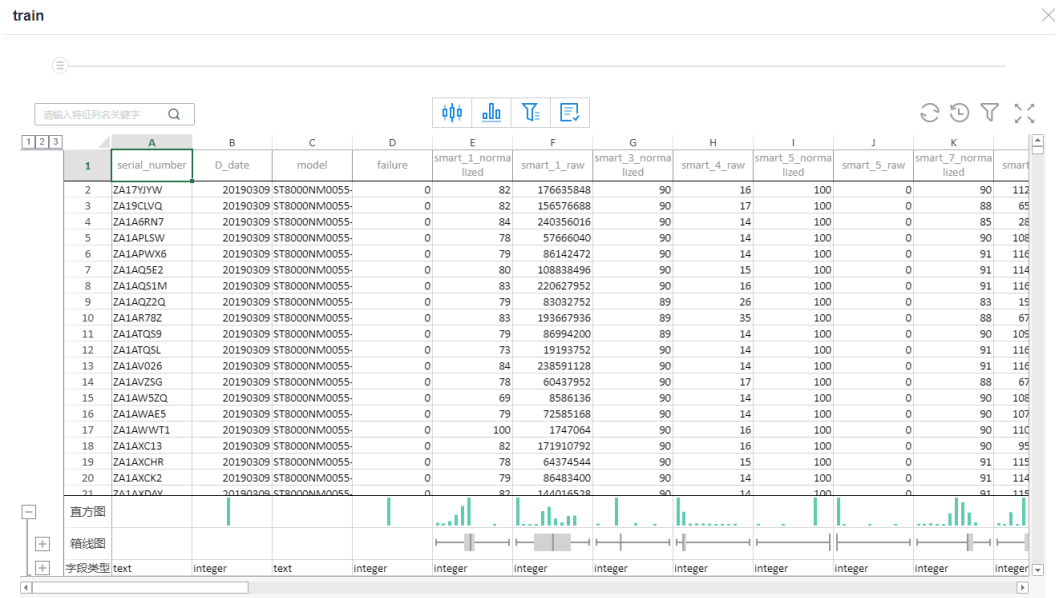
当前界面，支持如下操作：

- 在数据分析结果界面的“操作”列，单击  图标，可修改数据字段类型，目前数据类型可支持修改“TEXT”、“REAL”和“INTEGER”三种类型。
- 单击  图标，可设置当前字段为标签列。

图 4-12 数据分析结果

步骤7 单击界面右上角的“预览数据”，弹出预览数据界面，如图4-13所示。

图 4-13 预览数据界面



----结束

支持超大文件（10G）上传

支持多文件多目录上传，最多可上传10G大小。支持断点续传功能。

步骤1 在数据集界面，单击界面左上角的  图标。

弹出“导入数据”对话框，如图4-14所示。

参数说明如下所示：

- 数据集：从下拉框中选择已有数据集或编辑生成新数据集。示例“Case”。
- 数据类别：从下拉框中选择“多文件与目录（文件大小限制为10G）”。
- 实例名称：请根据实际情况配置。示例设置为“data”。
- 实例别名：请根据实际情况配置。支持设置为中文字。

图 4-14 导入数据

步骤2 单击“创建”，生成名称为“data”的数据。

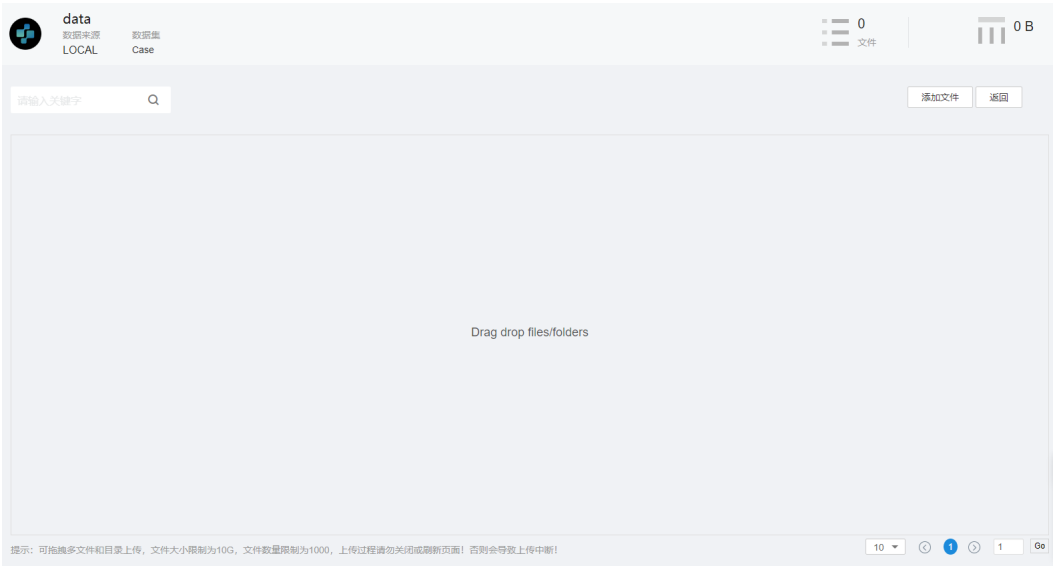
步骤3 在左侧数据集目录中，单击“data”，如图4-15所示。

右侧展示“data”的数据详情界面。

图 4-15 样例数据

步骤4 单击界面左上角的“上传”，进入文件拖拽上传面板界面，如图4-16所示。

图 4-16 文件上传面板



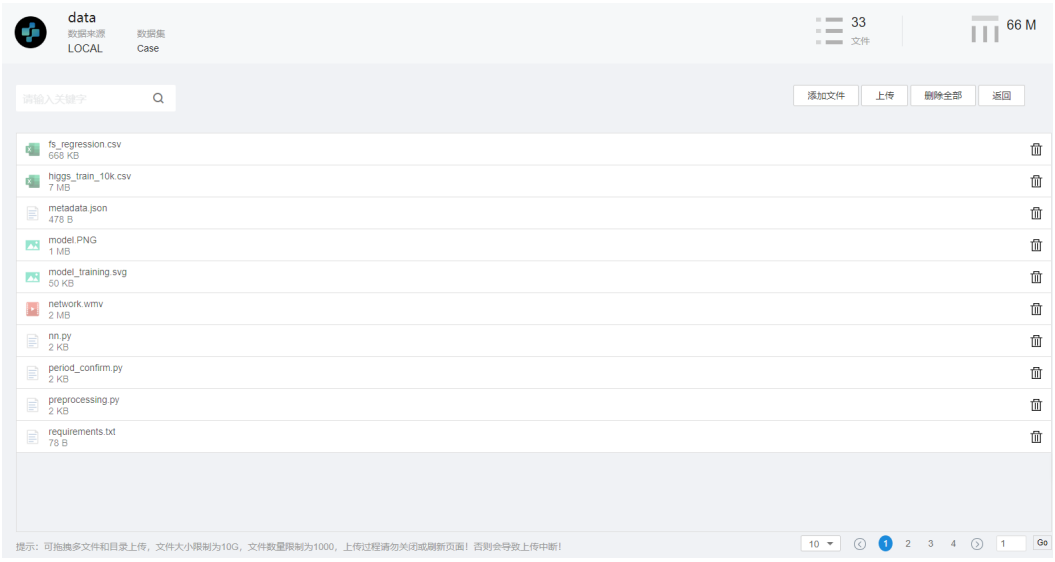
步骤5 从本地同时拖拽文件、数据文件目录到灰色边框展示区域，如图4-17所示。

目前支持的功能和限制如下所示：

- 当前右侧面板最多支持1000个文件，总大小最大为10G的上传任务。
- 文件上传过程中，请勿关闭或刷新页面，否则会导致数据上传中断。
- 大文件上传任务中断后，仍可从断点处继续上传。

  操作方式为：单击上传终端的文件右侧的 图标，从本地重新选择当前文件后，单击界面右上角的“上传”，完成断点续传。
- 支持删除或更新覆盖已上传的文件。

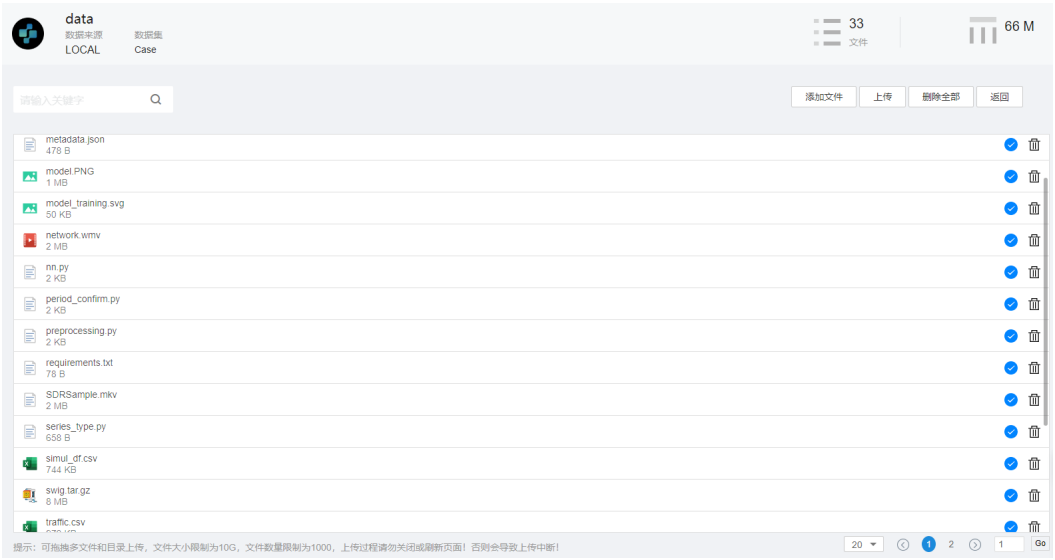
图 4-17 拖拽多文件和目录



步骤6 单击界面右上角的“上传”，等待数据上传完成，如图4-18所示。

批量上传本地文件时，支持按页分批上传文件。

图 4-18 上传数据



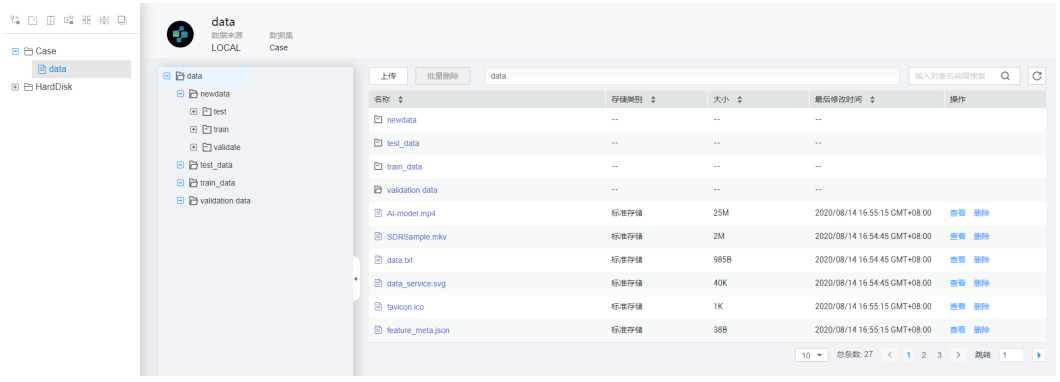
步骤7 等待数据上传完成后，单击左侧数据集目录中的“data”。

如图4-19所示，多文件数据集支持按目录结构进行树状展示。右侧文件列表支持分页展示，且支持对当前目录下面的文件进行前缀搜索（不支持模糊匹配）。

在右侧文件列表界面，单击具体数据文件右侧的“查看”，支持查看不同类型文件内容，包括：

- csv格式数据文件表格方式展示
- json文件格式化展示
- json文件、python等代码文件、markdown文件的CodeMirror渲染展示
- 绝大多数格式的图片文件
- mp3/ogg/wav格式的音频播放
- mp4/mkv/webm格式的视频播放

图 4-19 data 数据



----结束

4.5.3 数据集操作

对于数据样本量不足，或者在一定场景下，比如将采集的不同系统或网元的数据合并成一份数据的时候，用户可以在“数据集”界面中执行数据连接和数据联合操作：

- 数据连接：将特征列维度不完全相同的两份数据，合并成一份数据，用于扩展特征维度。
- 数据联合：将两份数据合并成一份数据，用于增加样本量。

数据连接

当用户需要基于已导入的数据实例，横向维度结合，扩展特征维度时，可以执行数据连接操作，新增一个数据实例。数据连接是基于主键字段列，采用leftouter、rightouter、inner、outer连接方式，连接两个数据实例。

说明

数据连接的两份数据的键值必须相同，否则系统无法进行数据连接。

将两份数据分别理解为左表和右表。连接方式说明如下：

- leftouter：以左表为主，返回所有左表数据以及匹配的右表数据。右表重复字段名加后缀__duplicate。
- rightouter：以右表为主，返回所有右表数据以及匹配的左表数据。左表重复字段名加后缀__duplicate。
- inner：以左表为主，返回左表和右表同时匹配的数据。右表重复字段名加后缀__duplicate。
- outer：以左表为主，返回左表和右表所有的数据。右表重复字段名加后缀__duplicate。

以如下表4-6、表4-7为例，键值为ID列，则按照leftouter、rightouter、inner、outer连接后的返回值分别如表4-8、表4-9、表4-10、表4-11。

表 4-6 左表数据

ID	Name	Height
1	A	1
3	B	2
5	C	2
7	D	2
9	E	2

表 4-7 右表数据

ID	Name	Weight
2	A	2
4	B	3
5	C	4
7	D	5

表 4-8 Leftouter 数据连接

ID	Name	Height	Name__duplicate	Weight
7	D	2	D	5
9	E	2	null	null
5	C	2	C	4
1	A	1	null	null
3	B	2	null	null

表 4-9 Rightouter 数据连接

ID	Name__duplicate	Height	Name	Weight
7	D	2	D	5
5	C	2	C	4
2	null	null	A	2
4	null	null	B	3

表 4-10 Inner 数据连接

ID	Name	Height	Name__duplicate	Weight
7	D	2	D	5
5	C	2	C	4

表 4-11 Outer 数据连接

ID	Name	Height	Name__duplicate	Weight
7	D	2	D	5
9	E	2	null	null
5	C	2	C	4
1	A	1	null	null
3	B	2	null	null
2	null	null	A	2

ID	Name	Height	Name__duplicate	Weight
4	null	null	B	3

数据连接操作步骤如下。

步骤1 单击数据目录区域框中的 ，弹出“数据连接”对话框，如图4-20所示。

图 4-20 数据连接界面

步骤2 配置“数据连接”对话框参数：

- 主数据集：主数据集、主数据实例、键值。
- 扩展数据集：扩展数据集、扩展数据实例、键值。
- 目标数据集：目标数据集、目标数据实例。其中目标数据名称只能以字母（A~Z、a~z）开头，由字母、数字（0~9）、下划线（_）、中划线（-）组成，不能以下划线或中划线结尾，且长度为[1-128]个字符。
- 连接方式：leftouter、rightouter、inner、outer。

步骤3 单击“确定”，执行数据连接。

数据连接完成后，系统在目标数据集下生成一个新数据实例，名称即为目标数据名称。

----结束

数据联合

如果说数据连接是数据的特征维度的扩展，那数据联合就是数据样本量的扩展。数据连接操作后，新生成的数据，其特征列会增多；数据联合操作后，数据实例的样本量会增多。

数据联合，是合并两份数据的样本，合并后数据的样本量是两份数据样本量的总和。

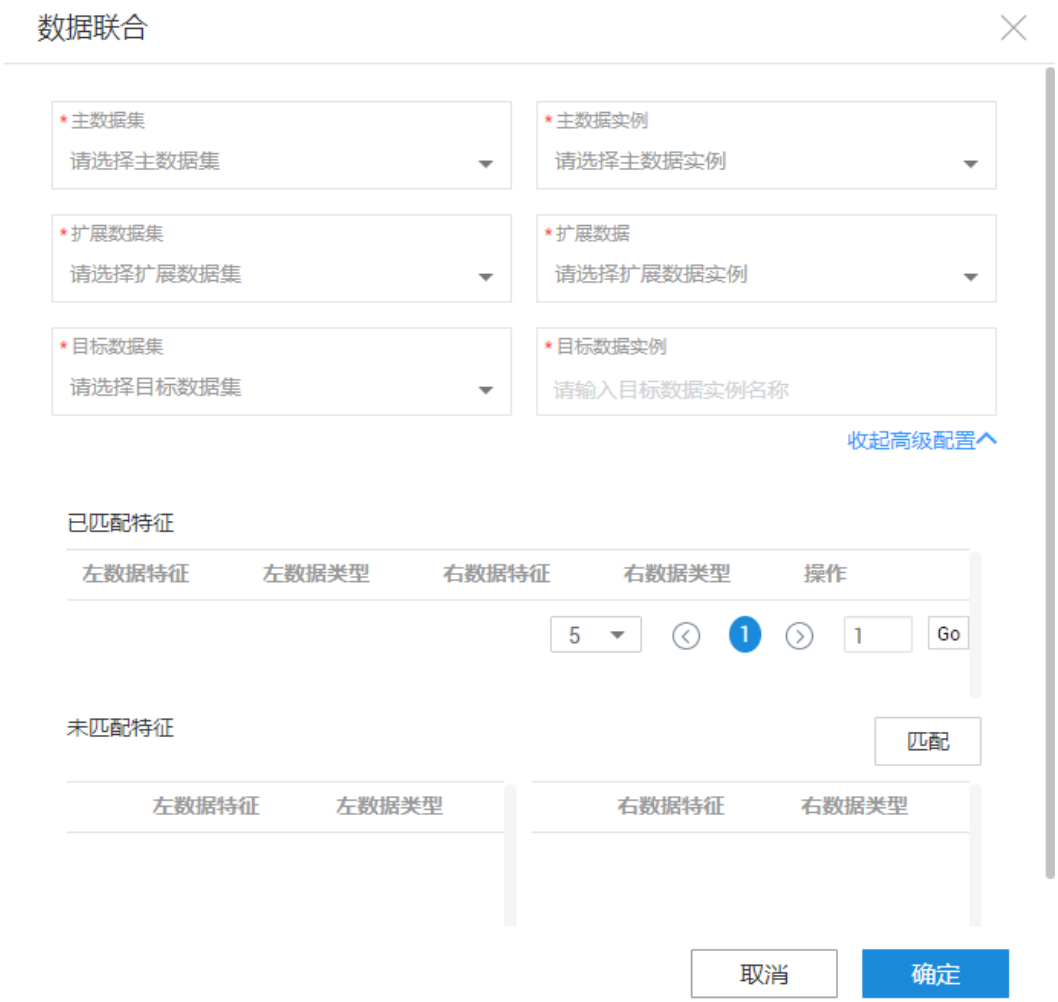
说明

左表和右表特征列数不一致时，按照如下情况处理：

- 左表特征列数多，右表不足的特征列补充空值。
- 右表特征列数多，以左表为准，删除右表多余的特征列。

步骤1 单击“数据目录”区域框中的，弹出“数据联合”对话框，如图4-21所示。

图 4-21 数据联合界面



步骤2 配置“数据联合”对话框参数：

- 主数据集：主数据集、主数据实例。

- 扩展数据集：扩展数据集、扩展数据实例。
- 目标数据集：目标数据集、目标数据实例。其中目标数据名称只能以字母（A~Z、a~z）开头，由字母、数字（0~9）、下划线（_）、中划线（-）组成，不能以下划线或中划线结尾，且长度为[1-128]个字符。

步骤3 展开高级配置，用户可以根据界面展示的左数据特征、左数据类型、右数据特征、右数据类型，手动配置需要匹配的特征列。

步骤4 单击“确定”，执行数据联合。

数据联合完成后，系统在目标数据集下生成一份新数据，名称即为目标数据实例名称。

----结束

4.6 特征工程

4.6.1 特征工程简介

用户可以通过特征工程对数据集进行数据处理、特征组合、特征转换等特征处理，最大限度的从原始数据中提取特征以供模型训练使用。训练平台的特征工程操作包括数据准备、特征操作和Notebook开发。此外，用户还可以将优质的特征工程发布成服务，以服务的形式对具备完全相同特征的数据进行预处理。

特征工程相关的基本概念：

- 特征工程：对数据进行特征处理操作的工程。
- 特征工程服务：将优质的特征工程发布成服务，用户可以直接调用该服务，对具备完全相同特征的数据进行特征处理。
- 特征工程任务：调用特征工程服务的过程。用户在调用特征工程服务的时候，需要基于特征工程服务新建任务。

特征工程管理页面

“特征工程”页面分为两个页签：特征处理工程和已发布服务。

- “特征处理工程”页签列出了已有的特征工程列表信息，如图4-22所示。在该页签，用户可以新建特征工程、编辑特征工程信息、导出特征工程、复制特征工程、删除特征工程，详情请参见表4-12。
- “已发布服务”页签列出了已发布的特征工程服务信息，如图4-23所示。在该页签，用户可以查看发布服务的详情，创建特征工程任务，删除特征工程服务，详情请参见表4-12。

图 4-22 特征处理工程页签

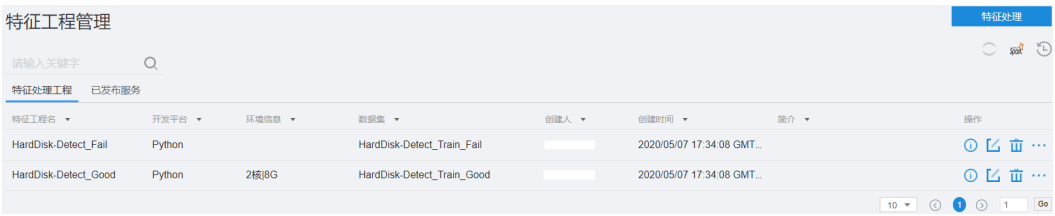


图 4-23 已发布服务页签

表 4-12 特征工程管理界面说明

页签	参数名称	参数说明
特征工程页面	特征处理	创建特征工程。
	请输入关键字	根据特征工程名称关键字，快速查找特征工程。
		单击图标，可查看Jupyterlab平台的环境信息，包括环境名称、状态、规格和剩余使用时间，停止运行环境的操作。
		Spark资源环境信息，用于数据集分析以及Spark特征工程。包括资源ID、状态、规格以及删除资源的操作。
		查看复制的特征工程的相关信息，包括任务类型、源特征工程、目标特征工程、创建时间和状态等信息。
特征工程页签	特征工程名	特征工程的名称。可以在创建特征工程时配置。
	开发平台	特征工程处理数据集的计算平台。 包括如下开发平台： • Jupyterlab • Python • Spark
	环境信息	包括运行环境的资源配置信息（“2核|8G”等）和运行状态（“创建中”、“运行中”等）。
	数据集	数据名称。
	创建人	创建特征工程的用户。
	创建时间	创建特征工程的时间。
	简介	特征工程的描述。
		进入特征工程操作界面。
		编辑特征工程相关信息，包括工程描述、AI引擎、规格等。
		删除特征工程。

页签	参数名称	参数说明
		单击操作列 ··· 图标后显示的下拉框中展示此图标。 下载特征工程包。
		单击操作列 ··· 图标后显示的下拉框中展示此图标。 复制生成新的特征工程。 训练平台支持将特征工程复制到项目公开组的其他项目中，对其他项目的数据进行特征处理。也支持复制到当前项目中，对其他数据进行特征处理。
		单击操作列 ··· 图标后显示的下拉框中展示此图标。 将特征工程发布成服务。详情请参见**发布服务**。 开发平台为“JupyterLab”的特征工程的操作列无此图标，此类特征工程的发布操作通过在JupyterLab环境编辑界面单击界面上方菜单栏中的发布图标完成，详情请参见**发布服务**。
已发布服务页签	服务名称	发布的特征工程服务名称。
	特征工程名	发布服务基于的特征工程名称。
	开发平台	特征工程处理数据集的计算平台。
	创建人	发布服务的用户名。
	创建时间	发布服务的时间。
	活动时间	最新执行特征工程任务的时间。
	简介	特征工程服务的简介。
		查看特征工程服务详情，包括特征工程任务的列表信息。
		创建特征工程任务。
		删除特征工程服务。

4.6.2 Python 和 Spark 开发平台

4.6.2.1 创建特征工程

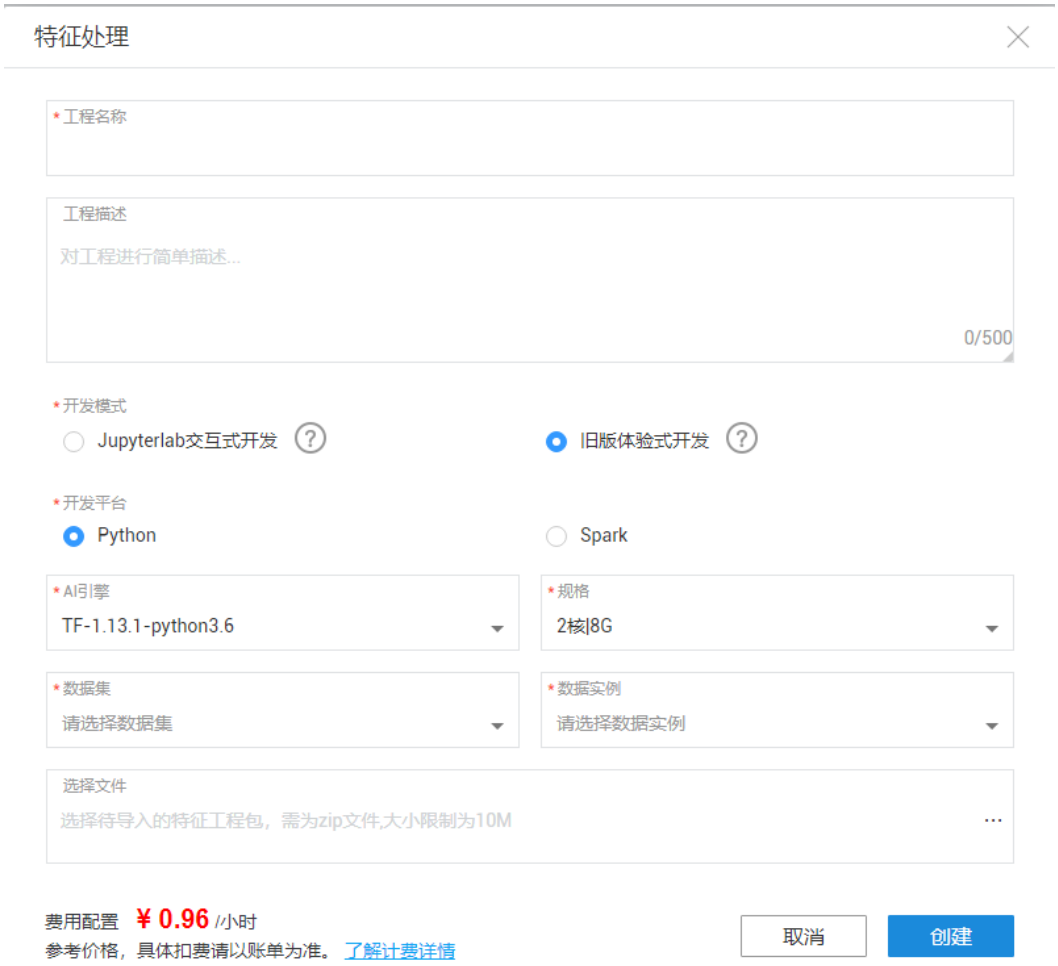
用户可以在“数据集详情”页面基于数据实例新建特征工程，对数据集执行特征操作；也可以在“特征工程管理”页面新建特征工程。我们以在“特征工程管理”页面创建特征工程为例，操作步骤如下。

步骤1 单击“特征工程管理”页面的 

。

弹出“特征处理”对话框。如图4-24所示。

图 4-24 创建特征工程



配置“特征处理”对话框参数，具体参见**特征工程参数配置说明**。

表 4-13 特征工程参数配置说明

参数名称	参数说明
工程名称	特征工程的名称。 只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线“_”、“-”组成，不能以下划线结尾，且长度为[1-50]个字符。
工程描述	特征工程描述信息。 最多不超过500个字符。

参数名称	参数说明
开发模式	特征工程的开发环境： • Jupyterlab交互式开发 基于JupyterLab的特征工程开发环境，具有良好的实时交互性，提供通用特征工程和数据分析的图形界面操作，以及用户自定义编码能力。适用于数据科学家，以及自定义算法场景。 • 旧版体验式开发 基于Web页面特征工程体验开发，适用于初学者及无码化特征工程。
开发平台	开发模式选择“旧版体验式开发”时展现，表示特征工程处理数据集的计算平台： • Python：对于小数据量的数据实例，选择使用Python。python分为local python与modelarts python，特征工程单步操作支持根据数据量大小，自动选择其中一种python执行，减少单步特征操作执行时间，提升用户体验。 • Spark：对于大数据量的数据实例，选择使用Spark，但是创建过程会比较慢。
AI引擎	特征处理算子运行平台。
规格	AI引擎的资源配置信息。
数据集	从下拉框中选择数据集。
数据实例	从下拉框中选择数据。
选择文件	直接导入已有的特征工程包，对数据进行特征处理。

步骤2 单击“创建”，创建特征工程，并进入“特征工程编辑”页面。如图4-25所示。

图 4-25 特征工程编辑界面

表 4-14 特征工程编辑界面说明

区域	说明
1	特征工程信息区域。包括开发平台、数据类型、数据集名称。
2	特征工程当前操作结果概览。包括当前数据行、原始数据行、当前数据列、原始数据列、当前执行的特征操作流个数。
3	包含如下操作： • 配置：配置“Notebook开发”中的超参。配置超参可以调用平台提供的SDK能力，以超参名称为“test”为例，SDK如下： `sai.get_hyper_param("test", type=str)` 用户单击“配置”，在弹出的“配置参数”对话框中分别输入“参数名”、“默认值”和“当前值”，即可修改超参值。 • 执行记录：查看全量数据应用的历史记录。并支持在“执行记录”中删除全量数据应用操作或重新执行全量数据应用操作。 • 执行：将特征操作流应用在导入特征工程的全量数据上，并生成经过特征处理的新数据。
4	特征操作明细区域。 单击“特征操作流总览”，查看特征操作流详情，单击每个特征操作名称前面的圆形图标，可以查看每个操作的特征处理效果。 仅支持对最后一个特征处理操作进行编辑修改或删除操作。
5	特征操作区域。支持数据采样、列筛选、数据准备、特征操作、Notebook开发、绘制Mini图、绘制图形、数据过滤以及数据验证功能。具体操作请参见“特征工程”各章节内容介绍。 • **数据采样** • **列筛选** • **数据准备** • **特征操作** • **Notebook开发** • 绘制Mini图：选中特征列，单击  图标，选择箱线图、折线图或面积图即可。支持同时选中多列进行操作。有些数据类型不支持绘制Mini图，如“Text”类型，操作时请注意界面右上角的提示信息。 • 绘制图形：选中特征列，单击  图标，选择需要展示的图形形式即可。支持同时选中多列进行操作。有些数据类型不支持绘制图形，如“Text”类型，操作时请注意界面右上角的提示信息。 • 数据过滤：类似Excel文档的数据过滤功能，同时支持对数据进行排序展示和有条件过滤展示。 • 数据验证：单击  图标，对全量数据进行数据验证，查看是否有空值，可通过单击  和  ，分别查看上一处和下一处空值。

----结束

4.6.2.2 数据采样

用户在执行特征操作前，可以先对数据进行采样。数据采样后，所有的特征操作都只对采样后的数据进行处理，可以减少特征操作处理的数据量，提升特征操作的处理速度。数据采样后，执行全量数据应用时，系统会将特征操作流应用在全量数据集上，生成经过特征处理后的新数据集，提供给模型训练使用。

说明

仅支持对刚导入的数据进行数据采样，不支持对已执行过特征操作的数据进行数据采样。

数据采样操作步骤如下。

步骤1 在特征工程首页，单击特征工程所在行，对应“操作”列的图标，进入特征操作界面。

步骤2 单击 数据采样 ，弹出“采样”对话框。

步骤3 配置采样参数如表4-15所示。

表 4-15 采样参数设置

参数名称	参数描述
采样方法	数据样本采样的方法。 包含如下方式： ● 随机采样：随机选取指定数量的样本。 ● 随机百分比：随机选取指定百分比的样本。 ● 前N条：按照从前往后的顺序选取指定数量的样本。 ● 全量：选取全部样本。
采样参数	采样方法为“随机采样”或“前N条”时，取值为记录数；采样方法为“随机百分比”时，取值为百分比。

步骤4 单击“确定”，系统提示“任务数据采样执行成功”，完成数据采样操作。

----结束

4.6.2.3 列筛选

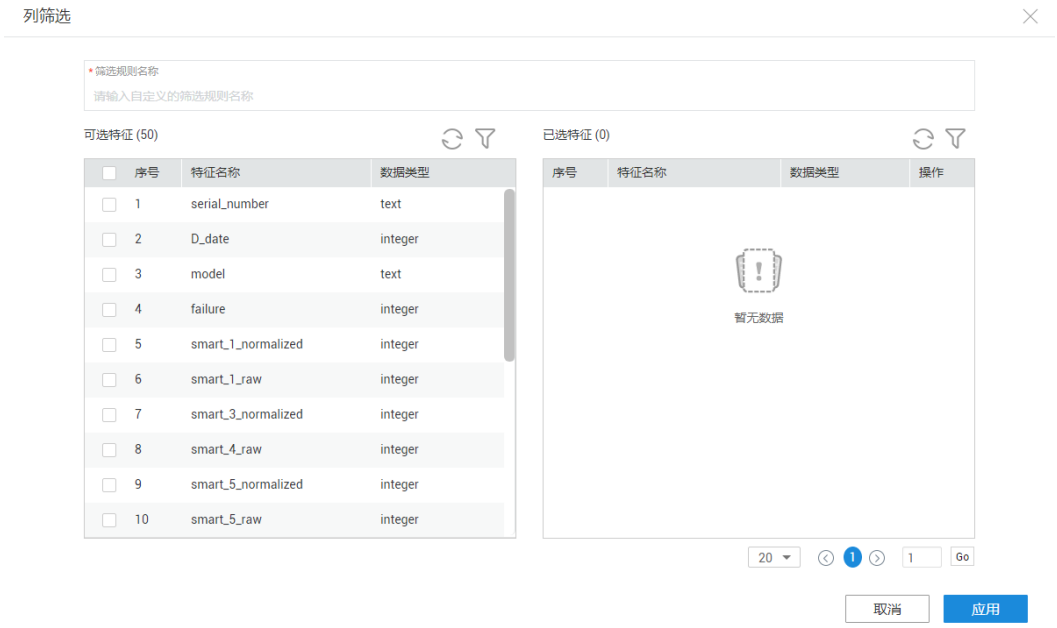
“列筛选”筛选的是特征列。如果用户需要重点查看分析特定特征列，可以通过列筛选完成。

列筛选操作步骤如下。

列筛选

步骤1 单击特征工程操作界面的 ，弹出“列筛选”对话框，如图4-26所示。

图 4-26 列筛选



其中，界面说明如下所示：

- 筛选规则名称：为即将设置的筛选规则设置名称。

  筛选成功后，在特征工程操作界面可以单击 图标查看筛选历史，筛选记录内的规则名即为此处设置的筛选规则名称，单击筛选历史记录内的筛选规则名可以查看对应的筛选结果。
- 可选特征：展示当前数据的所有特征信息。
- 已选特征：展示用户在“可选特征”中勾选出的所有特征，支持删除已选特征。

步骤2 在“可选特征”框中勾选需要显示的特征列。

同时包含下述操作：

- 单击 图标，通过设置“列区间”、“列关键词”、“数据类型”、“数据质量”，快速查找特征列，如图4-27所示。

图 4-27 特征筛选条件

- 单击 图标，可恢复图4-27中的筛选条件为默认设置。

步骤3 在“已选特征”框中再次确认并删除不需要显示的特征列。

如果用户需要进一步筛选出不需要显示的特征列，可以通过操作列的 取消显示。

用户可以单击“已选特征”框中的 ，快速搜索不需要显示的特征列，并根据搜索结果，通过 取消显示：

- 通过配置“列关键词”搜索仅显示包含该关键词的特征列。
- 通过选择“数据类型”仅显示该数据类型的特征列。

步骤4 单击“应用”，完成特征筛选。

----结束

查看筛选历史

单击特征工程操作界面的 图标，弹框中自动展示已执行过的所有列筛选操作记录。支持单击某条记录，查看列筛选的执行结果。

重置筛选条件

单击特征操作界面的 图标，回退列筛选操作。

4.6.2.4 数据准备

数据集中的数据导入特征工程后，可能存在空值、冗余、数据不足等情况，或者用户需要将多次导入的数据实例进行数据联合。以上情况，都可以在数据准备中进行操

作。当前数据准备包含的功能有：数据修复、数据过滤、数据联合、数据连接、数据去噪。

数据修复

用户可以在数据修复中对单列进行空值修复、无效值修复，以及根据取值范围进行修复，多列或者全选所有特征列进行空值修复。系统有默认的修复策略，用户也可以自行配置修复策略。操作步骤如下。

步骤1 单击表头，选择需要进行数据修复的特征列。

步骤2 单击“数据准备”，从下拉框中选择“数据修复”。

弹出“数据修复”对话框。参数设置如下所示：

- 检查“已选择特征”是否为用户选择的特征列。
- 配置“修复策略”如表4-16所示。

表 4-16 修复策略配置

参数	参数说明
NA值	对特征列样本中的空值进行修复，修复策略有： – 丢弃：直接丢弃空值所在行。 – 替换值：以用户指定的数值替换空值。 系统默认丢弃空值。
无效值	对特征列样本中的无效值进行修复，修复策略有： – 丢弃：直接丢弃无效值所在行。 – 替换值：以用户指定的数值替换无效值。 系统默认丢弃无效值。
取值范围	对特征列样本中指定取值范围内的数据进行修复。 用户配置样本数据的取值范围，系统丢弃取值范围之外的数据。 系统默认不根据取值范围进行数据修复。

步骤3 单击“确定”，执行数据修复。

----结束

数据过滤

用户可以配置单列特征的过滤方式和过滤规则，筛选掉冗余的样本数据行，或者仅保留有效的样本数据行。操作步骤如下。

步骤1 单击表头，选择需要进行数据过滤的特征列。

步骤2 单击“数据准备”，从下拉框中选择“数据过滤”。

弹出“数据过滤”对话框。参数设置如下所示：

- 检查“已选择特征”是否为用户选择的特征列。
- 配置“过滤方式”和“过滤规则”如表4-17所示。

表 4-17 过滤策略配置

参数	参数说明
过滤方式	过滤方式有两种： – 保留行：保留符合过滤规则的样本数据行。 – 丢弃行：丢弃符合过滤规则的样本数据行。
过滤规则	过滤规则根据样本数据值进行配置： – 大于：保留或丢弃大于指定值的样本数据行。 – 小于：保留或丢弃小于指定值的样本数据行。 – 等于：保留或丢弃等于指定值的样本数据行。

步骤3 单击“确定”，执行数据过滤。

----结束

数据联合

特征工程数据联合的原理与数据集中数据联合的原理相同。具体请参见**数据联合**。特征工程以当前打开的特征工程的数据实例为左表，“数据联合”对话框中数据集的数据为右表。

数据联合操作步骤如下。

步骤1 单击“数据准备”，从下拉框中选择“数据联合”。

弹出“数据联合”对话框。参数设置如下所示：

- 在“数据集”、“数据实例”对应的下拉框中选择需要联合的数据集和数据实例。

  系统会先将当前特征工程的数据实例和设置的数据实例进行自动匹配，并在“数据实例”框下方展示匹配结果。
- 展开高级配置，用户可以在“已匹配特征”栏下查看系统自动匹配的特征记录。在“未匹配特征”栏下，用户可以根据界面展示的左表数据特征、左表数据类型、右表数据特征、右表数据类型，手动配置需要匹配的特征列，不同数据类型的特征无法匹配。如需取消匹配，可单击记录操作列的“取消匹配”。

步骤2 单击“确定”，执行数据联合。

----结束

数据连接

特征工程数据连接的原理与数据集中数据连接的原理相同，具体请参见**数据连接**。特征工程的数据连接参数说明如下：

- 当前打开的特征工程的数据实例为左表，“数据连接”对话框中数据集的数据为右表。

- 主键为左表的键值，外键为右表的键值。主键和外键必须相同。
- 连接方式为leftouter、rightouter、inner、outer，与数据集中数据连接相同。

数据连接操作步骤如下。

步骤1 单击表头，选中一列数据作为连接的参考列。

步骤2 单击“数据准备”，从下拉框中选择“数据连接”。

弹出“数据连接”对话框。参数设置如下所示：

- “数据集”、“数据实例”对应的下拉框中选择需要连接的数据集和数据集版本作为右表。
- 在“主键”下拉框中选择主键作为左表的ID，在“外键”下拉框中选择外键作为右表的ID。主键和外键必须相同。
- 在“连接方式”下拉框中选择连接方式。

步骤3 单击“确定”，执行数据连接。

----结束

数据去噪

用户可以通过数据去噪，筛选掉时间序列中的异常数据。噪声分析方法：

1. 通过局部线性回归的方法对数据进行平滑处理，得到每个点对应的预测值。
2. 通过观测值与预测值之间的误差error的3sigma确定误差上限，超出上限的点为噪声点。

系统会从原始数据中去除上述噪声点，并采用线性插值的方法对去除噪声的数据进行填充。操作步骤如下。

步骤1 单击表头，选择需要数据去噪的特征列。

步骤2 单击“数据准备”，从下拉框中选择“数据去噪”。

弹出“数据去噪”对话框。检查“已选择特征”是否为用户选择的特征列。

步骤3 单击“确定”，执行数据去噪。

----结束

4.6.2.5 特征操作

特征操作主要是对特征的样本数据值进行修改，也可以重命名、删除、筛选特征列。同时训练平台集成了基于开源的交互式开发调试工具，支持用户编辑算法自定义修改特征列。训练平台支持的特征操作有重命名、归一化、数值化、标准化、特征离散化、One-hot编码、数据变换、删除列、选择特征、卡方检验、信息熵、新增特征、PCA。

重命名

训练平台特征工程支持用户修改特征名，操作步骤如下。

步骤1 单击表头，选中需要执行重命名的一个特征列。

不支持同时选择多列进行重命名操作。

步骤2 单击“特征操作”，从下拉框中选择“重命名”。

弹出“重命名”对话框。参数设置如下所示：

- 检查“已选择特征”是否为用户选择的特征列。
- 新特征名：新特征名不能与数据集中其他特征列的名称重复，且只能由字母、数字（0~9）、下划线（_）和（-）组成。

步骤3 单击“确定”。

在“特征操作流总览”区域会新增一个“重命名”节点。

----结束

归一化

归一化是一种简化计算的方式。训练平台支持三种归一化算法：

- MaxAbsScaler：将特征列的样本数据映射到[-1,1]区间上。
- MinMaxScaler：将特征列的样本数据映射到[0,1]区间上。
- StandardScaler：处理后的样本数据服从均值为0，方差为1的标准正态分布。

归一化操作过程如下。

步骤1 单击表头，选中需要执行归一化的特征列。

步骤2 单击“特征操作”，从下拉框中选择“归一化”。

弹出“归一化”对话框。参数配置如下所示：

- 检查“已选择特征”是否为用户选择的特征列。
- 配置“归一化算法”。

步骤3 单击“确定”。

在“特征操作流总览”区域会新增一个“归一化”节点。

----结束

数值化

很多情况下样本数据并不是数值型，例如“性别”的值为男和女，“姓名”的值为“Alex”。这时无法执行特征操作，需要通过数值化将其转换为数值型。数值化的思路是根据特征列的样本数据的种类进行编码，数值化后样本数据为取值范围在[0,样本数据种类-1]区间内的整型数据。以特征列Sepal（样本数据为abcadc）为例，数值化后，样本数据为012032。

步骤1 选中需要执行数值化的特征列。

步骤2 单击“特征操作”，从下拉框中选择“数值化”。

弹出“数值化”对话框。检查“已选择特征”是否为用户选择的特征列。

步骤3 单击“确定”。

在“特征操作流总览”区域会新增一个“数值化”节点。

----结束

标准化

标准化支持L1_norm和L2_norm两种算法对特征列的样本数据进行处理：

- L1_norm：所有样本数据的绝对值求和作为分母；样本数据作为分子。将样本数据映射到（ -1,1 ）区间。
- L2_norm：所有样本数据求平方和后开根号作为分母；样本数据作为分子。将样本数据映射到（ -1,1 ）区间。

标准化操作步骤如下。

步骤1 单击表头，选中需要执行标准化的特征列。

选中的特征列必须为数值型。

步骤2 单击“特征操作”，从下拉框中选择“标准化”。

弹出“标准化”对话框。参数配置如下所示：

- 检查“已选择特征”是否为用户选择的特征列。
- 配置“标准化算法”。

步骤3 单击“确定”。

在“特征操作流总览”区域会新增一个“标准化”节点。

----结束

特征离散化

特征离散化是将特征列连续的样本数据离散化为[0，离散数量-1]区间内的整型数据。

特征离散化操作步骤如下。

步骤1 单击表头，选中需要执行特征离散化的特征列。

选中的特征列必须为数值型。

步骤2 单击“特征操作”，从下拉框中选择“特征离散化”。

弹出“特征离散化”对话框。参数配置如下所示：

- 检查“已选择特征”是否为用户选择的特征列。
- 配置“离散数量”。

步骤3 单击“确定”。

在“特征操作流总览”区域会新增一个“特征离散化”节点。

----结束

One-hot 编码

One-hot编码定义是使用N位状态寄存器来对N个状态进行编码。直观来说，在特征工程中One-hot是将特征列根据样本数据的种类拆分成多列，将原特征列数据映射到新特征列中，样本数据相同编码为1，不同则编码为0。以特征列Sepal样本数据为（ 2,9,2,8,4 ）为例，One-hot编码后，则拆分为四列特征列，每列样本数据为：

- Sepal_2：10100
- Sepal_4：00001
- Sepal_8：00010
- Sepal_9：01000

One-hot编码操作步骤如下。

步骤1 单击表头，选中需要执行One-hot编码的特征列。

选中列不同值的数量不能小于2，不能大于100。

步骤2 单击“特征操作”，从下拉框中选择“One-hot编码”。

弹出“One-hot编码”对话框。检查“已选择特征”是否为用户选择的特征列。

步骤3 单击“确定”。

在“特征操作流总览”区域会新增一个“One-hot编码”节点。

----结束

数据变换

数据变换是通过以自然常数e为底的自然对数（log）、以自然常数e为底的指数函数（exp）对特征列的样本数据进行变换：

- log：如果当前样本数据比较大，可以通过对数函数进行变换。
- exp：如果当前样本数据比较小，可以通过指数函数进行变换。

数据变换操作步骤如下。

步骤1 单击表头，选中需要执行数据变换的一个特征列。

选中的特征列必须为数值型，且不支持同时选中多列进行数据变换。

步骤2 单击“特征操作”，从下拉框中选择“数据变换”。

弹出“数据变换”对话框。参数配置如下所示：

- 检查“已选择特征”是否为用户选择的特征列。
- 配置“算法”。

步骤3 单击“确定”。

在“特征操作流总览”区域会新增一个“数据变换”节点。

----结束

删除列

特征操作支持删除数据集中指定的特征列，操作步骤如下。

步骤1 单击表头，选中需要执行删除的特征列。

步骤2 单击“特征操作”，从下拉框中选择“删除列”。

弹出“删除列”对话框。检查“已选择特征”是否为用户选择的特征列。

步骤3 单击“确定”。

在“特征操作流总览”区域会新增一个“删除列”节点。

----结束

选择特征

特征操作支持选择并保留数据集中指定的特征列，删除其余特征列。操作步骤如下。

步骤1 单击表头，选中需要执行的特征列。

步骤2 单击“特征操作”，从下拉框中选择“选择特征”。

弹出“选择特征”对话框。检查“已选择特征”是否为用户选择的特征列。

步骤3 单击“确定”。

在“特征操作流总览”区域会新增一个“选择特征”节点。

----结束

卡方检验

卡方检验通过计算数据集的特征列和标签列之间的偏离程度（即卡方值）筛选出有价值的特征列。将卡方值由小到大排序，筛选出TOPN的特征列：

- 特征列与标签列之间的偏离程度越大，卡方值越大，说明特征列与标签列不符
- 特征列与标签列之间的偏离程度越小，卡方值越小，说明特征列越接近于标签列
- 如果特征列与标签列完全相等，卡方值为0，说明特征列与标签列完全符合

以投掷硬币为例，投掷一枚硬币50次，记录正面特征值和反面特征值的实际值分别是多少。假设硬币是均匀的，正面特征值的理论值是25，反面特征值的理论值也是25，如果实际投掷结果为：正面22，反面28，则卡方值为（22-25）*（22-25）/25+（28-25）*（28-25）/25=0.72。

说明

- 选定特征列不同值的数量不能超过10000。
- 如果特征列的样本数据中存在负数，在进行卡方检验之前，系统会采用MinMaxScaler算法对特征列进行归一化。
- 如果特征列的样本数据为字符型，在进行卡方检验之前，系统会先对特征列进行数值化，再采用MinMaxScaler算法进行归一化。

卡方检验操作方法如下。

步骤1 单击表头，选中一个特征列作为标签列。

步骤2 单击“特征操作”，从下拉框中选择“卡方检验”。

弹出“卡方检验”对话框。参数设置如下所示：

- 检查“已选择特征”是否为用户选择的特征列。
- 配置“变换特征数”，保留指定“变换特征数”的特征列。

步骤3 单击“确定”，执行卡方检验。

在“特征操作流总览”区域会新增一个“卡方校验”节点。

----结束

信息熵

信息熵是通过计算数据集的特征列与标签列之间的相关性筛选出有价值的特征列。相关性越大，信息熵越大；相关性越小，信息熵越小。将信息熵由大到小排序，筛选出信息熵较大的有价值的特征列。

信息熵操作方法如下。

步骤1 单击表头，选中一个特征列作为标签列。

选定列不同值数量不能超过100。

步骤2 单击“特征操作”，从下拉框中选择“信息熵”。

弹出“信息熵”对话框。参数设置如下所示：

- 检查“已选择特征”是否为用户选择的特征列。
- 配置“变换特征数”，保留指定“变换特征数”的特征列。

步骤3 单击“确定”，执行信息熵。

在“特征操作流总览”区域会新增一个“信息熵”节点。

----结束

新增特征

新增特征支持用户基于已有的特征列，按照样本数据行的维度，通过求和、求均值，构造出新的特征列。例如，两个特征列ID1（2,7,1）和特征列ID2（3,2,7），求和后构造出的特征列为ID_SUM（5,9,8）。

说明

选择的多列特征必须是数值型，并且没有异常值。

新增特征操作步骤如下。

步骤1 单击表头，依次选中多个特征列。

步骤2 单击“特征操作”，从下拉框中选择“新增特征”。

弹出“新增特征”对话框。参数设置如下所示：

- 检查“已选择特征”是否为用户选择的特征列。
- 配置“新增列名”。列名只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线（_）、（-）组成，不能以下划线结尾，且长度为[1-50]个字符。
- 配置“新增规则”：
  - sum：按照样本数据行的维度，对已选择特征列进行求和。
  - mean：按照样本数据行的维度，对已选特征列的求均值。

步骤3 单击“确定”，执行新增特征。

----结束

PCA

PCA的实质就是在尽可能代表原特征的情况下，将原特征进行线性变换，寻找数据分布的最优子空间，从而达到降维、去相关的目的。

训练平台支持两种主成分分析算法：

- PCA：主成分分析。将数据集从高维投影到低维，从而用极少的几个特征来涵盖大部分的数据集信息。主成分分析认为，沿某特征分布的数据的方差越大，则该特征所包含的信息越多，也就是所谓的主成分。适用于线性可分的数据集。
- KPCA：基于核函数的主成分分析。KPCA与PCA基本原理相同，只是需要先升维再进行投影，因为有些非线性可分的数据集只有在升维的视角下才线性可分。

说明

在执行PCA之前，系统会对所有数值型的特征字段先做标准化处理。对于字段类型为text的字段，系统会先做数值化处理，然后做标准化处理。

PCA操作步骤如下。

步骤1 单击“特征操作”，从下拉框中选择“PCA”。

弹出“PCA”对话框。参数配置如下所示：

- 转换数目：转换后的特征列数。例如，待降维的特征列有5列，配置转换数目为2后，执行PCA后，系统会计算出2个涵盖信息最多的两个特征列。
- 选择算法：PCA和KPCA。Spark开发平台不支持KPCA算法。

步骤2 单击“确定”，执行PCA。

----结束

4.6.2.6 Notebook 开发

用户在Notebook开发环境中编写算法，自定义修改特征列。操作步骤如下。

步骤1 单击“Notebook开发”。

弹出“Notebook开发”对话框，如图4-28所示，进行操作名称及操作描述填写。

图 4-28 Notebook 开发界面

步骤2 单击“确定”。

进入“特征工程算法编辑”界面。Notebook算法开发界面同模型训练算法界面，详情请参见**编辑代码**。用户可以编辑算法文件“**.py”，按“Ctrl+S”保存算法。

说明

用户编辑完成通过notebook开发的特征处理算法后，一定要单击界面右上角的“保存”，防止当前编辑的算法内容全部丢失。

步骤3 （可选）配置Notebook调试环境，调试算法。

1. 单击 

，弹出Notebook配置对话框。

   如果有已经创建好的Notebook环境，直接选中“运行中”的环境，单击“保存”即可。否则需要重新创建Notebook开发环境，操作步骤如下：

   a. 分别选择AI引擎和规格，单击“创建Notebook环境”。
   b. 系统自动创建环境，待环境状态为“运行中”时，选中该环境，单击“保存”。
2. 单击“*.ipynb”文件进入算法调试界面。
3. 单击 

，调试算法。

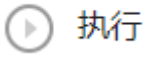
步骤4 算法调试成功后，单击界面右上角的“保存”，开始运行自定义特征处理算法。

在“特征操作流总览”区域会新增一个“自定义操作”节点。

----结束

4.6.2.7 全量数据应用

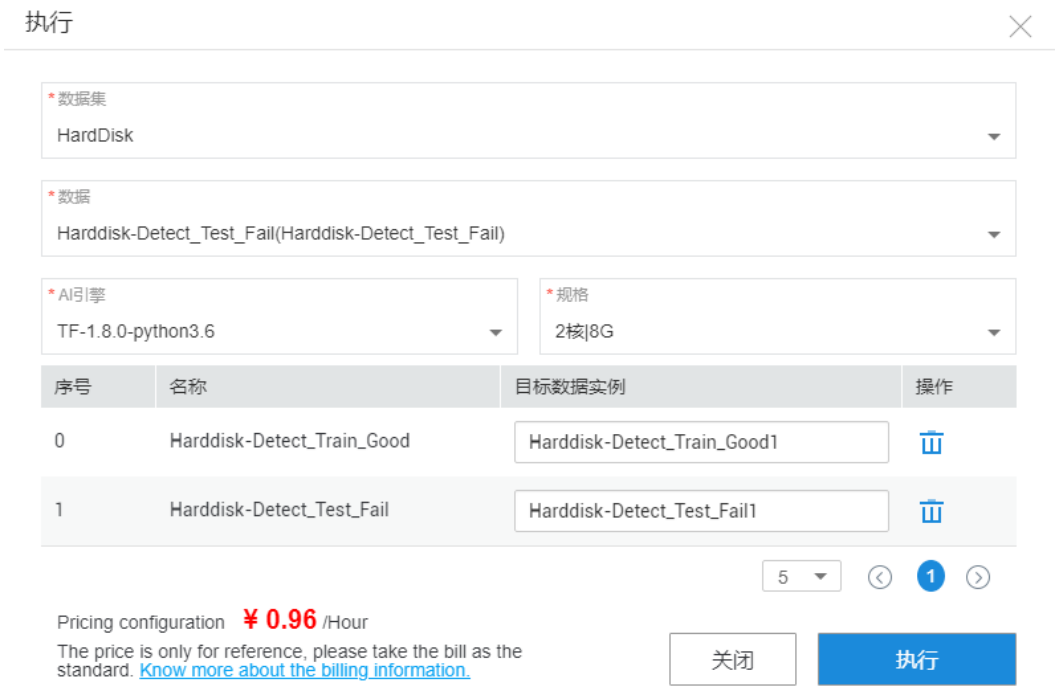
特征操作完成后，需要单击“执行”，应用特征操作流至全量数据。同时支持选择其他数据集和数据实例，应用当前特征操作流。全量数据应用操作步骤如下。

步骤1 单击 

。

弹出“执行”对话框。如**图4-29**所示。

图 4-29 全量数据应用



步骤2 在“数据集”和“数据”下拉框中，分别选择数据集和数据。

支持同时添加多份数据，每份数据必须满足与当前特征工程中的数据特征维度完全相同。

其中，“目标数据实例”为特征处理后生成的数据实例名称，请根据实际情况配置。

步骤3 单击“执行”，对数据执行特征操作流。

系统自动生成经过特征处理后的数据，支持用户在“数据集”中查看。

用户可以执行下述操作：

- 在特征工程详情页面单击“执行记录”，查看数据实例名称、目标数据实例名称、时间、状态。其中“操作”列，支持重新执行全量数据应用操作、基于新生成的数据实例创建算法，或删除新生成的数据实例操作。
- 在数据集页面查看应用特征操作流后生成的新数据实例，“数据集”中此类数据的数据来源为“FEATURE”。

----结束

4.6.2.8 发布服务

如果当前特征工程操作流处理效果比较好，可以得到比较优质的训练数据，可以将当前的特征工程发布成服务。复用此特征工程服务对其他数据进行相同的特征操作。

发布特征工程服务

步骤1 在特征工程首页的“特征工程”页签，单击特征工程对应“操作”列的 ···，在展开的下拉框中单击 图标。

弹出“发布服务”对话框。配置对话框参数：

- 服务名称：特征工程服务名称。只能以字母（A~Z a~z）开头，由字母、数字（0~9）、中划线（-）组成，不能以中划线结尾。
- 服务描述：特征工程服务描述信息。字数不能超过256。

步骤2 单击“确定”，发布特征工程服务。

步骤3 单击“已发布服务”页签，可以查看发布的特征工程服务信息，包括服务名称、特征工程名、开发平台、创建人、创建时间、活动时间和简介等信息。用户还可以执行以下操作：

- 查看特征工程服务的详情，如图4-30所示。特征工程服务界面介绍如表4-18所示。

图 4-30 特征工程服务界面

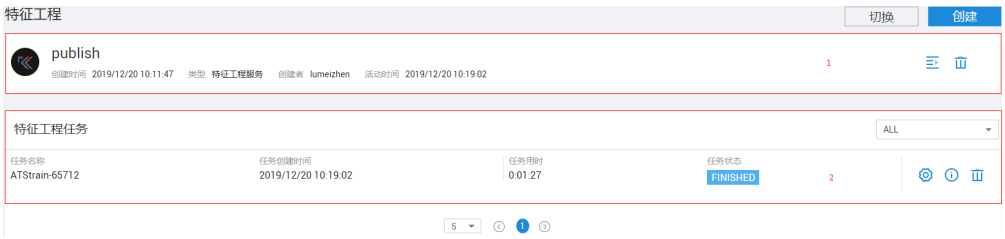
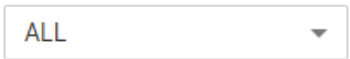


表 4-18 特征工程服务界面说明

区域	参数名称	参数说明
1（特征工程服务）	创建时间	特征工程服务创建时间
	类型	特征工程服务
	创建者	创建特征工程服务的用户
	活动时间	最近一次特征工程任务执行的时间
		创建新的特征工程任务，详细请参考**创建特征工程任务**
		删除特征工程服务
2（特征工程任务）		根据状态快速检索验证任务
	任务名称	特征工程任务的名称
	任务创建时间	特征工程任务创建的时间
	任务用时	特征工程任务的执行时长
	任务状态	特征工程任务的执行状态

区域	参数名称	参数说明
		查看特征工程任务的超参配置
		查看特征工程任务的运行日志
		删除特征工程任务

- 基于特征工程服务创建特征工程任务，请参见**创建特征工程任务**。
- 删除特征工程服务。

用户也可以在特征工程服务页面，单击右上角的“创建”，基于其他特征工程新建特征工程服务。

----结束

创建特征工程任务

可以在“已发布服务”页签，单击特征工程服务对应“操作”列的  创建特征工程任务，也可以在已发布特征工程服务详情界面，单击右上角的  创建特征工程任务。本文以在“特征工程服务”界面操作为例，操作步骤如下。

步骤1 在“特征工程服务”界面，单击右上角的  ，弹出“创建任务”对话框。

步骤2 配置对话框参数，如**表4-19**所示。

表 4-19 创建特征工程任务

区域	参数名称	参数描述
任务信息	任务名称	特征工程任务的名称。 名称只能以字母（A~Z a~z）开头、由字母、数字（0~9）、下划线（_）组成，不能以下划线结尾，长度范围为[1,26]。
	数据集	在下拉框中选择要执行特征工程任务的数据集。
	数据实例	在下拉框中选择要执行特征工程任务的数据。
	目标数据	特征工程任务执行完成后，系统会在数据集中自动生成来源为FEATURE的新目标数据。 只能以字母（A~Z a~z）开头，由字母、数据（0~9）、下划线（_）（-）组成，不能以下划线或中划线结尾，且长度为[1-128]个字符。
环境配置	AI引擎	特征工程算子的开发平台。

区域	参数名称	参数描述
	规格	AI引擎的资源配置信息。

步骤3 单击“创建”，新建特征工程任务。

任务执行过程中，可以单击 查看运行日志。

任务执行结束后，可以在数据集中查看数据来源为“JOB”的新目标数据。

----结束

4.6.3 JupyterLab 开发平台

4.6.3.1 创建特征工程

用户可以在“数据集详情”页面基于数据实例新建特征工程，对数据集执行特征操作；也可以在“特征工程管理”页面新建特征工程。我们以在“特征工程管理”页面创建特征工程为例，操作步骤如下。

步骤1 单击“特征工程管理”页面的 特征处理 。

弹出“特征处理”对话框。如图4-31所示。

图 4-31 创建特征工程

配置“特征处理”对话框参数，具体参见表4-20。

表 4-20 特征工程参数配置说明

参数名称	参数说明
工程名称	特征工程的名称。 只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线“_”、“-”组成，不能以下划线结尾，且长度为[1-50]个字符。
工程描述	特征工程描述信息。 最多不超过500个字符。
开发平台	特征工程处理数据集的计算平台JupyterLab。
规格	计算平台的资源配置信息，请根据实际情况选择。
实例	创建JupyterLab运行环境的实例。可以从下拉框中选择已创建的运行环境或选择“新建一个新环境”。

步骤2 单击“创建”。

在特征工程首页“特征工程”页签默认生成一行新的特征工程。

等待特征工程“环境信息”列状态由“创建中”变更为“运行中”，即JupyterLab环境实例创建完成。

步骤3 单击特征工程所在行，对应“操作”列的图标。

进入JupyterLab环境编辑界面。

步骤4 在弹出的选择内核的弹窗中，选择内核版本，单击“Select”。

进入JupyterLab环境编辑界面，如图4-32所示。JupyterLab环境编辑界面说明如表4-21所示。

图 4-32 JupyterLab 环境编辑界面

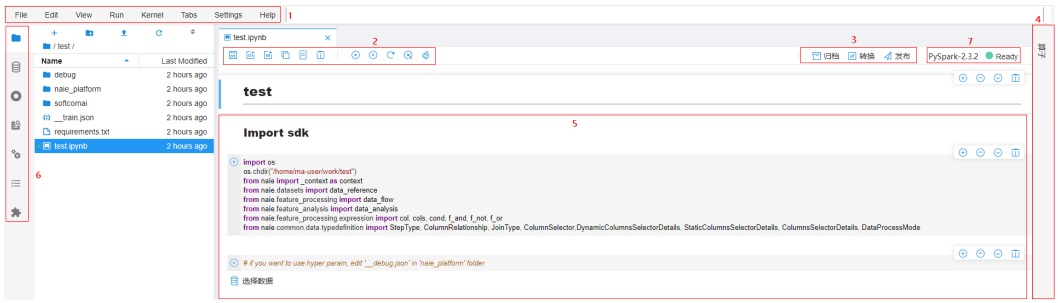


表 4-21 JupyterLab 环境编辑界面说明

区域	说明
1	JupyterLab平台自带的菜单项。
2	文件编辑时的快捷操作按钮。

区域	说明
3	训练平台预置的特征服务发布、基于整个Jupyterlab的模型包归档以及转换特征工程操作主文件格式的能力。
4	JupyterLab预置的算子，主要包含数据处理、模型训练以及迁移学习能力。
5	特征工程操作编辑区域。特征工程操作的主文件为后缀名称是“ipynb”的文件。
6	特征工程详细信息查看区域。 下述为图标的详细释义： ●  ：特征工程目录列表。展示与特征工程同名的所有目录。双击目录名称，可查看特征工程包含的所有子目录和文件，如下所示： – softcomai目录：训练服务提供的SDK。 – _train.json：超参配置信息。 – *.ipynb：特征工程代码编辑和调试运行主文件。 – requirements.txt：训练服务第三方依赖包列表。用户可根据实际需要写入依赖的第三方包。示例：tensorflow==1.8.1。 ●  ：数据集目录列表。展示用户当前项目OBS空间中的所有数据集列表。双击数据集目录，可查看其包含的数据列表。 ●  ：查看运行的所有JupyterLab环境信息，可单击“SHUT DOWN”，停止运行环境。 ●  ：Jupyterlab功能集。 ●  ：属性检查器，可查看右侧编辑界面上各Cell的属性。 ●  ：展示所有基于JupyterLab平台创建的特征工程操作流。单击特征操作名称，可直接定位到特征操作在编辑界面的位置。 ●  ：Jupyterlab第三方拓展功能管理。
7	特征工程的内核信息，单击当前内核版本可重新选择内核。

----结束

4.6.3.2 数据集

平台提供的SDK能力，用户可以通过如下两种方式获取解释：

- 通过新增代码框，输入“?dataflow.rename_columns”的形式，运行代码框，查看释义。
- 通过界面右上角“帮助中心”中的“SDK文档”，查看SDK文档中释义。

选择数据

用户在执行特征操作前，需要先选择数据。

可以任选下述一种方式选择数据：

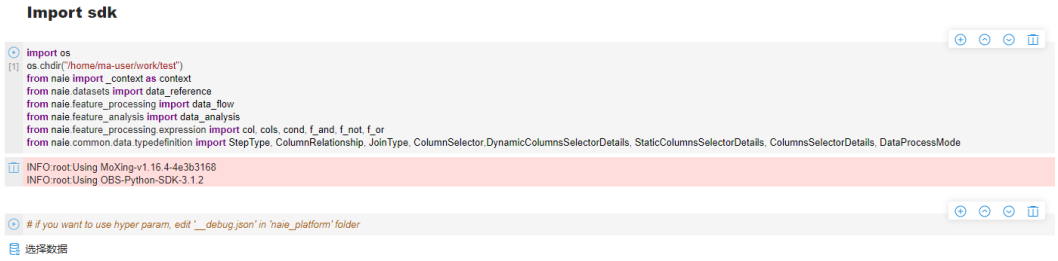
- 在编辑界面，单击“Import sdk”下方的“选择数据”。
- 在特征工程右上角，选择“算子 > 数据处理 > 数据集 > 选择数据”。

选择数据操作步骤如下。

步骤1 单击如图4-33所示的  图标，运行“Import sdk”内容。

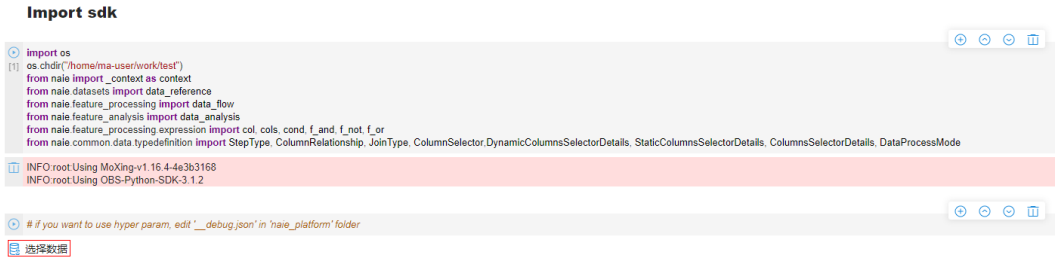
“Import sdk”必须放在所有操作的最前面执行，否则执行“选择数据”会报错。

图 4-33 导入 SDK



步骤2 在编辑界面，单击“选择数据”，如图4-34所示。

图 4-34 选择数据



步骤3 在如图4-35所示的框中输入待处理的“数据集”和“数据集实例”。

如果数据是通过本地上传的方式，且本地上传时，“数据类别”参数设置为“多文件与目录（文件大小限制为10G）”，则需要同时设置“数据文件列表”和“数据文件格式”，将本地上传的多目录和文件同时添加进来，系统会自动进行数据集合并。

图 4-35 设置选择的数据集及实例

右侧参数说明如表4-22所示。

表 4-22 选择数据

参数	参数说明
数据集	从下拉框中选择已有的数据集。
数据集实例	从下拉框中选择数据实例。
是否为时序数据	选择数据为时序数据时，可开启此开关。 开启开关后，需要同时配置如下参数： ● 时间列：输入时间字段名称。 ● 时间格式：时间字段的时间格式。默认为“自动解析”，系统会自动解析时间格式。 ● ID列：数据集的标识列。
数据文件列表	当数据通过本地上传，且本地上传时，“数据类别”参数设置为“多文件与目录（文件大小限制为10G）”，则需要同时设置“数据文件列表”和“数据文件格式”，将本地上传的多目录和文件同时添加进来，系统会自动进行数据集合并。各文件的列明需要完全相同。
数据文件格式	本地上传文件的格式，请根据实际情况选择。
是否检测周期与平稳性	开启开关会检测时序数据的周期，或判断指定的周期是否为时序数据的周期，以及检测时序数据是否平稳。 如果开启此开关，运行时间会较长，默认关闭此开关。

参数	参数说明
数据引用变量名	当特征工程需要选择多份数据，如进行AutoML模型训练时，如果需要同时配置训练数据、测试数据、集成学习数据以及验证数据，则需要在选择数据时选择多份数据，使用此参数给每份选定的数据命名，以免产生冲突。
操作流变量名	特征工程中存在多个操作流时，使用此参数分别为操作流对象命名，以免产生冲突。

步骤4 单击图标，运行“选择数据”代码框内容。数据实例绑定成功。

----结束

生成数据实例

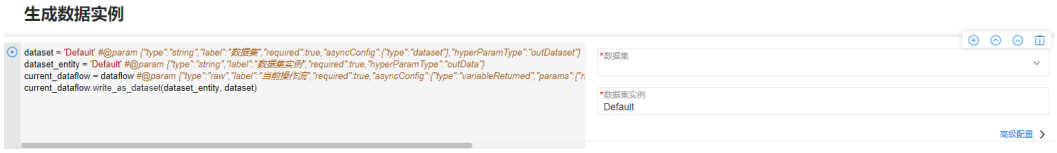
在特征工程编辑界面完成所有特征操作后，需要将特征操作流应用于选择的数据，并生成经过特征处理的新数据。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据集 > 生成数据实例”。

在编辑界面，展示如图4-36所示的内容。

图 4-36 生成数据实例



右侧参数说明如表4-23所示。

表 4-23 生成数据实例参数说明

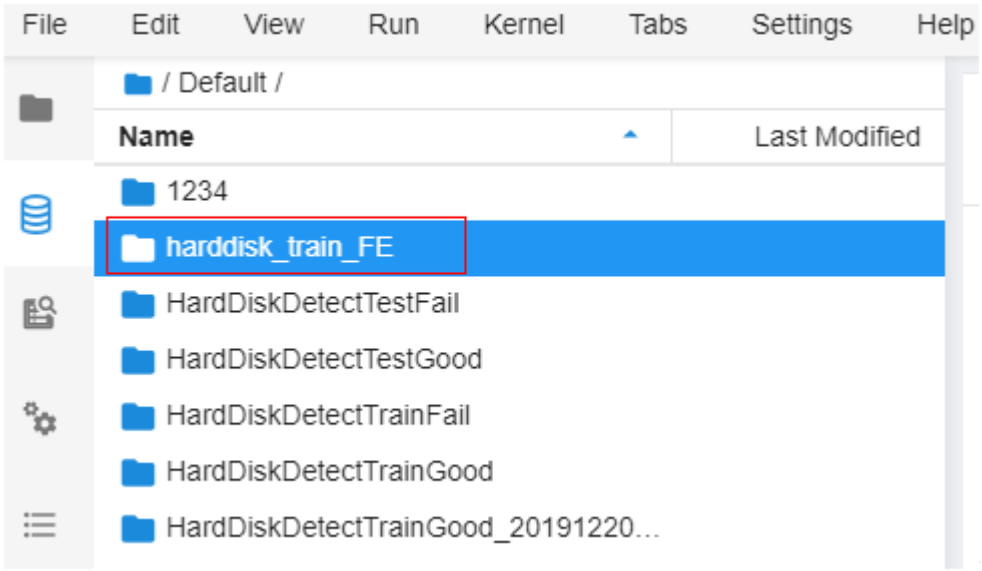
参数	参数说明
数据集	从下拉框中选择已有的数据集。
数据集实例	经过特征处理后新生成的数据名称，支持用户自定义。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。

步骤2 单击图标，在已有数据集下面生成新的数据。

可通过双击如图4-37所示的数据集名称目录，打开查看新生成的数据集。

支持双击数据集名称，在右侧界面查看数据内容。

图 4-37 数据集列表



----结束

4.6.3.3 数据探索

选择数据后，可以对选定的数据进行探索操作，包括数据统计、图表分析、特征分析以及时序分析。

数据统计

支持对当前特征数据进行全量展示，包括所有特征字段对应的字段类型，字段值。还支持对某个特征字段进行统计，统计类型包括平均值、方差、最大值、最小值、百分比分位数。支持特征列绘制直方图、箱线图、折线图及面积图。

操作步骤如下所示。

步骤1 在特征工程操作编辑区域，“选择数据”代码框下方单击“数据探索”。

数据探索文件在特征工程编辑界面上默认靠右展示，左侧为特征工程代码编辑主文件（*.ipynb）展示区域。用户可以通过按鼠标拖拽的方式设置数据探索文件的展示区域。左键单击数据探索文件标题区域，长按鼠标拖拽数据探索文件待出现蓝色底纹区域框后松开鼠标即可。数据探索文件可和特征工程代码编辑主文件（*.ipynb）分上下、左右区域同时展示，也可同级展示。同级展示时，特征工程编辑界面只展示一个文件界面，通过单击文件标题切换文件界面。

步骤2 展开“数据统计”页签，查看特征数据全量表。

步骤3 单击数值类型的特征列列名，可查看该特征列绘制的直方图、箱线图等，如果数据为时序数据，可绘制趋势图。在全量特征统计表下方可查看该特征列的数值统计。

时序数据的数值统计中，时间列信息的“时间间隔是否均匀”为“否”时，需要执行时序数据预处理操作。

如果时序数据在选择数据操作时，“是否检测周期与平稳性”开关关闭，则数据的“是否平稳”和“周期（样本数）”统计项可手动进行检测操作，数据量较大时，检测执行时间会较长，用户可自行选择是否检测。

----结束

图表分析

支持对当前特征数据进行图表展示。

操作步骤如下所示。

步骤1 在特征工程操作编辑区域，“选择数据”代码框下方单击“数据探索”。

步骤2 展开“图表分析”页签，按需求设置图表图形。界面参数说明如表4-24所示。

表 4-24 图表分析界面参数明

功能入口	功能说明	参数	参数说明
	图表类型及图表展示参数设置。	图表类型	特征数据可展示的图表类型，包括散点图、折线图、直方图、箱线图、散点图矩阵、KDE曲线、3D散点图。 如果特征数据为时序数据，支持的图表类型分别有趋势图、直方图、箱线图、KDE曲线、ACF与PACF。
		标题	特征数据图表标题。
		X轴	单击“…”，从特征数据的特征列中选择数据列作为图表X轴。
		Y轴	单击“…”，从特征数据的特征列中选择数据列作为图表Y轴。
		Z轴	单击“…”，从特征数据的特征列中选择数据列作为图表Z轴。
		列名	单击“…”，选定特征数据的特征列作为直方图、箱线图、KDE曲线、散点图矩阵、ACF与PACF展示的数据来源。
		视觉维度配置	“是否启用视觉维度”为开启状态时，单击“标签列名”对应的“…”，选定特征数据的特征列作为散点图、折线图、3D散点图的视觉维度标签，视觉维度标签将展示在图表右上角。
		包含高斯分布曲线	是否展示高斯分布曲线开关。图表类型为直方图时展示。
		直方图柱数	直方图展示柱的数量。 图表类型为直方图时展示。

功能入口	功能说明	参数	参数说明
		Lag	绘制ACF与PACF图表时设置的滞后阶数。
	图表外观设置	主题	图表主题
		散点图设置	设置散点图标记点类型和标记点大小。
		折线图设置	设置折线图线条是否平滑、标记点类型以及大小。
		视觉维度设置	设置视觉维度的样式，如颜色、大小、形状等。
	截取及清空图表展示图		截取当前图表图形，截取后的图形展示在左侧空白区域。
			清空截取的图表图形。

步骤3 单击右下方“保存至特征工程”可将绘制的图表保存至特征工程操作编辑区域。

----结束

特征分析（特征选择）

特征选择就是使用算法对特征进行相关性分析，根据结果从众多特征中剔除不重要的特性，从而保留重要的特性。

当前系统支持如下两种特征选择方法：

- 过滤法(Filter)

  按照发散性或者相关性对各个特征进行评分，设定待选择评分数最高的特征个数，选择特征。
- 包装法(Wrapper)

  算法每次根据皮尔逊相关系数选择一个相关系数最大的特征进行丢弃，并进行模型训练得出精度，当精度低于设置的阈值时，停止丢弃特征。

使用过滤法时提供如下算法：

- 卡方检验

  卡方检验就是统计样本的实际观测值与理论推断值之间的偏离程度。实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，如果卡方值越大，二者偏差程度越大；反之，二者偏差越小；若卡方值为0，表明实际值与理论值完全符合。
- F检验

  F检验是一种在零假设之下，统计值服从F-分布的检验。
- 信息增益

  信息增益是对两个随机变量之间相关信息量的度量，值越大说明变量之间的相关性越强。

上述算法中，卡方检验、F检验和信息增益可用于分类任务，F检验和信息增益可用于回归任务。

步骤1 在特征工程操作编辑区域，“选择数据”代码框下方单击“数据探索”。

步骤2 选择“特征分析”页签。

步骤3 在左侧目录树上单击“特征选择”。

步骤4 设置“标签列”、“方法”、“算法”等参数，具体参数说明如表4-25所示。

表 4-25 特征选择界面参数说明

参数	参数说明
标签列	单击“…”选择标签列，用以分析特征列和标签列的相关性。
方法	特征分析可选用的方法，目前支持如下两种方法： ● 过滤法(Filter) ● 包装法(Wrapper)
算法	“方法”选择“过滤法”时可选用的具体分析算法，目前支持如下算法： ● 卡方检验 ● F检验 ● 信息增益
选择特征数	特征分析完成后按相关性大小展示的Top N特征数。
标签列是否为类别型	标签列设置后，该参数会根据标签列的类型自动判断是否为类别型，用户可使用默认值。
随机种子	“算法”为“信息增益”时设置，用以生成随机数。
排除特征列	执行包装法前需要排除的特征列，这些被排除的列不参与后续特征选择，单击“…”选择排除特征列。
指标阈值	模型训练精度阈值。使用“包装法(Wrapper)”会对特征进行反复训练，当训练结果精度低于设置的阈值时，停止丢弃特征。
分析	提交分析请求。
停止	提交分析任务至分析完成期间，可单击此按钮终止分析任务。
创建Snapshot	截取Top N柱状图。用户可以另存图片至本地使用。
清空	清空界面上的相关性分析Top N柱状图截图。

步骤5 单击“分析”。

说明

系统自动分析完成后，将以柱状图和列表形式展示分析结果，柱状图中展示的特征列的个数即为设置的“选择特征数”值。列表默认按照相关性评分降序展示所有的特征列。

步骤6 选择特征列。

- 保留分析结果所有Top N个特征列。
  a. 单击Top N柱状图结果下方的“应用”。
     页面跳转至特征工程操作编辑区域并生成“选择特征”代码框，“列选择”下展示的“列名”为柱状图展示的所有特征列。
  b. 单击  图标，运行“选择特征”代码框内容。
- 保留部分分析结果中的特征列。
  a. 勾选“分析结果”列表中特征列前的复选框，如需选择所有特征列，可勾选表头中的复选框。
  b. 单击“分析结果”列表下方的“应用”。
     页面跳转至特征工程操作编辑区域并生成“选择特征”代码框，“列选择”下展示的“列名”为用户勾选的特征列。
  c. 单击  图标，运行“选择特征”代码框内容。

----结束

特征分析（ACE）

ACE（Alternating Conditional Expectation）是一种在回归分析中寻找响应变量Y（标签）与预测变量X（特征）之间最佳转换的算法，这些（转换后的）预测变量和（转换后的）响应变量之间产生最大的线性效应。ACE分析只支持回归类任务。

步骤1 在特征工程操作编辑区域，“选择数据”代码框下方单击“数据探索”。

步骤2 选择“特征分析”页签。

步骤3 在左侧目录树上单击“ACE”。

步骤4 设置“标签列”、“列名”、“特征列变换初始化方法”等参数，具体参数说明如表4-26所示。

表 4-26 ACE 界面参数说明

参数	参数说明
标签列	响应变量，单击“…”选择标签列，仅支持单列选择。
列名	预测变量，单击“…”选择列名，支持多列选择。

参数	参数说明
特征列变换初始化方法	ACE分析时，特征列的初始化方式，支持如下特征列变换初始化方法： ● zeros 表示0作为初始值。 ● zero-mean 表示将特征值减去均值后的值作为初始值。 ● std 表示将特征值减去均值再除以方差后的值作为初始值。
标签列变换初始化方法	ACE分析时，标签列的初始化方式，支持如下标签列变换初始化方法： ● zero-mean ● std
迭代误差容忍度	迭代终止条件，当迭代误差达到“迭代误差容忍度”值时，终止迭代。默认值为“0.001”。
最大迭代次数	迭代终止条件，当迭代次数达到“最大迭代次数”时，终止迭代。默认值为“100”。 “迭代误差容忍度”和“最大迭代次数”无论哪个先满足，迭代都会终止。
近邻样本数	算法迭代过程中，需要求解到每个点的近邻数量，默认值为“100”。
是否使用kd-tree	是否使用k-维树来搜索近邻数。k-维树是一种分割k维数据空间的数据结构。
分析	提交分析请求。
停止	提交分析任务至分析完成期间，可单击此按钮终止分析任务。
创建Snapshot	截取ACE分析图。用户可以另存图片至本地使用。
清空	清空界面上ACE分析截图。

步骤5 单击“分析”。

分析完成后右侧展示分析结果图，可单击“保存至特征工程”将分析结果图保存到特征工程操作编辑区域。

----结束

时序分解

时间序列的变化会受到长期趋势（T）、季节变动（S）、周期变动（C）以及不规则变动（L）的影响，时序数据分解是指使用加法模型或乘法模型将原始数据拆分成上述四部分。

步骤1 在特征工程操作编辑区域，“选择数据”代码框下方单击“数据探索”。

步骤2 选择“时序分析”页签。

步骤3 在左侧目录树上单击“时序分解”。

步骤4 设置“时间列”、“特征列”、“模型”等参数，具体参数说明如表4-27所示。

表 4-27 时序分解界面参数说明

参数	参数说明
时间列	待分解时序数据的时间列。
特征列	待分解时序数据特征列。
模型	时序数据分解使用的分解模型，支持： • 加法模型 如果季节变动的幅度以及趋势和周期的波动都不随时间变化而变化，则比较适合使用加法模型。 • 乘法模型 如果季节变动的幅度或趋势和周期的波动随时间变化而变化，则比较适合使用乘法法模型。
周期	时序数据周期值。
分析	提交分析请求。
停止	提交分析任务至分析完成期间，可单击此按钮终止分析任务。

步骤5 单击“分析”。

分析完成后右侧展示分析结果图，可单击“保存至特征工程”将分析结果图保存到特征工程操作编辑区域。

----结束

异常检测

时序数据序列中存在模式不一致的异常点（如时序数据超出正常范围的上/下界，突然的上升或下降，趋势改变），时序数据的异常检测旨在快速准确地找到这些异常点。

步骤1 在特征工程操作编辑区域，“选择数据”代码框下方单击“数据探索”。

步骤2 选择“时序分析”页签。

步骤3 在左侧目录树上单击“异常检测”。

步骤4 设置“时间列”、“特征列”、“异常类型”等参数，具体参数说明如表4-28所示。

表 4-28 异常检测界面参数说明

参数	参数说明
时间列	待异常检测时序数据的时间列。
特征列	待异常检测时序数据的特征列。
异常类型	异常检测类型: • 数值范围 表示检测平稳时序数据是否异常，给出异常判断参考区间。 • 突升/突降 表示检测平稳时序数据中突增或突降的异常点。
异常区间获取方法	获取用于判断时序数据异常的上/下界区间的方法，支持: • 箱线图 • 3 Sigma • 两者任意一个检测到异常 • 两者同时检测到异常
突变点个数	“异常类型”为“突升/突降”时展示，表示需要检测到平稳时序数据中突增或突降点的个数。 默认值为5，检测结果有可能会小于这个个数。
是否进行周期分解	“异常类型”为“突升/突降”时展示，表示如果待检测数据为周期数据，是否需要进行周期分解，用于增强数据的差异性。 默认关闭。
一个周期内的数量值	“是否进行周期分解”开启时展示此参数，表示进行周期分解时，一个周期内的数据量。
是否进行过滤	“异常类型”为“突升/突降”时展示，表示是否对检测出的Top N个点进行二次过滤。 默认关闭。
过滤阀值	“是否进行过滤”开启时展示，表示如果对检测出的Top N个点进行二次过滤，则该参数作为过滤阈值，小于阈值的点将被认为是突变点。

参数	参数说明
分析	提交分析请求。
停止	提交分析任务至分析完成期间，可单击此按钮终止分析任务。

步骤5 单击“分析”。

分析完成后右侧展示分析结果图，可单击“保存至特征工程”将分析结果图保存到特征工程操作编辑区域。

----结束

4.6.3.4 特征数据采样

如果数据量太大，造成特征操作等待的时间长，用户可以通过采样功能减少特征处理的数据量，提升特征处理的速度。

数据采样提供如下两种方式，请根据实际情况进行选择：

- 随机采样：按照比例进行样本数据的随机采样。
- 分层采样：如果一个特征或多个特征组合样本值的类型多样，为保证采样数据的多样性，可以对不同类型的数据分别设置采样比例。

数据采样有如下两个入口：

- 特征工程“算子”菜单栏中，选择“数据处理 > 数据采样”。下文采样步骤使用此入口进行描述。
- 特征工程操作编辑区的“随机采样”、“分层采样”快捷入口。

随机采样

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据采样 > 随机采样”。

界面新增如图4-38所示的内容。

图 4-38 随机采样

参数含义如表4-29所示：

表 4-29 随机采样参数说明

参数	参数说明
采样比例	数据采样比例，取值范围(0,1)。请根据实际情况设置。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“随机采样”代码框内容。

----结束

分层采样

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据采样 > 分层采样”。

界面新增如图4-39所示的内容。

图 4-39 分层采样

参数含义如表4-30所示。

表 4-30 分层采样参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	列筛选方式为“列选择”时展示，请根据实际情况，单击“...”设置单列或者多列。
正则表达式	列筛选方式为“正则匹配”时展示，请根据实际情况输入正则表达式设置特征列。

参数	参数说明
fractions	为不同数据类型的样本数据，分别设置采样比例。 示例：{(0,): 0.2, (1,): 0.8}，其中(0,)和(1,)分别为特征列的组合样本数据。
seed	改变随机数生成器生成随机数的种子。取值必须为整数。 默认值为空，即不对分层采样产生影响。seed值不固定的时候，每次采样出来的样本数量，以及每层采的哪些行都是不固定的。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“分层采样”代码框内容。

----结束

4.6.3.5 特征数据清洗

去除空值

如果特征列中存在空值，“去除空值”操作可以去除掉空值所在行的样本数据。

去除空值有如下两个入口：

- 特征工程“算子”菜单栏中，选择“数据处理 > 数据清洗 > 去除空值”。下文去除空值步骤使用此入口进行描述。
- 特征工程操作编辑区的“去除空值”快捷入口。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据清洗 > 去除空值”。

界面新增如图4-40所示的内容。

图 4-40 去除空值

参数含义如表4-31所示。

表 4-31 去除空值参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	列筛选方式为“列选择”时展示，请根据实际情况，单击“…”设置单列或者多列。
正则表达式	列筛选方式为“正则匹配”时展示，请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
列关系	去除空值和特征列的关系。 取值如下所示： ● all：如果一行数据，满足设置列中的所有特征列均为空值，则丢弃此行数据。 ● any：如果一行数据，满足设置列中的任一特征列有空值，则丢弃此行数据。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“去除空值”代码框内容。

----结束

空值填充

如果样本数据量较少，或者用户可以根据特征等信息推断出实际的样本值，则可通过“空值填充”操作，替换空值。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据清洗 > 空值填充”。

界面新增如图4-41所示的内容。

图 4-41 空值填充

参数含义如表4-32所示。

表 4-32 空值填充参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	列筛选方式为“列选择”时展示，请根据实际情况，单击“…”设置单列或者多列。
正则表达式	列筛选方式为“正则匹配”时展示，请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
填充为	空值替换后的数据。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“空值填充”代码框内容。

----结束

数据替换

如果特征列中的数据有误或者与用户的心理预期不符，用户可以通过“数据替换”批量替换错误的数据。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据清洗 > 数据替换”。

界面新增如图4-42所示的内容。

图 4-42 数据替换

参数含义如表4-33所示。

表 4-33 数据替换参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	列筛选方式为“列选择”时展示，请根据实际情况，单击“…”设置单列或者多列。
正则表达式	列筛选方式为“正则匹配”时展示，请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
被替换值	需要替换的数据。
替换为	替换后的数据。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“数据替换”代码框内容。

----结束

数据映射

将特征列中的数据映射替换为用户需要的数据后，生成一个新的特征列。原有特征列不受影响，仍然保留。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据清洗 > 数据映射”。

界面新增如图4-43所示的内容。

图 4-43 数据映射

参数含义如表4-34所示。

表 4-34 数据映射参数说明

参数	参数说明
列名	请根据实际情况，单击“…”设置待映射特征列。仅支持设置单列。
新列名	输入经过数据映射后新生成的特征列名称。
被替换值	需要替换的数据。
替换为	替换后的数据。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“数据映射”代码框内容。

----结束

数据过滤

如果提供的数据存在一定的误差，比如只能为正数的特征，存在一部分负值，可通过“数据过滤”的方式将负值所在行都丢弃掉。

操作步骤如下所示。

数据过滤有如下两个入口：

- 特征工程“算子”菜单栏中，选择“数据处理 > 数据清洗 > 数据过滤”。
- 特征工程操作编辑区的“数据过滤”快捷入口。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据清洗 > 数据过滤”。

界面新增如图4-44所示的内容。

图 4-44 数据过滤

参数含义如表4-35所示。

表 4-35 数据过滤参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	列筛选方式设置为“列选择”时才会展示。 通过单击“…”图标，在弹出的对话框中，选择一个或者多个特征列。
正则表达式	列筛选方式设置为“正则匹配”时才会展示。 请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
表达式	数据过滤的表达式。 如果对单列数据进行过滤，可使用符号（>，>=，<，<=，==）进行过滤。示例如下所示，其余依次类推。 ● 取大于0的数据：col(columns[0]) > 0 ● 取等于2的数据：col(columns[0]) == 2 如果对多列数据进行过滤，可使用符号（f_and，f_or，f_not等符号）进行过滤。示例如下所示，其余依次类推。 ● 取两列值相等的数据：(col(columns[0])) == (col(columns[1])) ● 取两列值均是2的数据：f_and((col(columns[0]) == 2), (col(columns[1]) == 2))
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击图标，运行“数据过滤”代码框内容。

----结束

去重

如果特征列中存在重复的数据，可通过“去重”操作，删除数据重复的样本行。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据清洗 > 去重”。

界面新增如图4-45所示的内容。

图 4-45 去重

参数含义如表4-36所示

表 4-36 去重参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	列筛选方式为“列选择”时展示，请根据实际情况，单击“...”设置单列或者多列。
正则表达式	列筛选方式为“正则匹配”时展示，请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“去重”代码框内容。

----结束

4.6.3.6 特征数据合并

数据连接

数据连接是将特征列维度不完全相同的数据集连接成一份数据。数据集特征不完全相同的原因，比如现网中不同 系统采集的数据。其原理与“数据集”界面的数据连接原理相同，具体请参见**数据连接**。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据合并 > 数据连接”。

界面新增如图4-46所示的内容。

图 4-46 数据连接

参数含义如表4-37所示。

表 4-37 数据连接参数说明

参数	参数说明
右数据	当前特征工程绑定的数据为左数据，需要输入进行数据连接的右数据。
主键	左数据和右数据通过“主键”进行数据匹配。单击“…”设置主键。
连接方式	数据连接的方式。 包含如下选项： • left：返回所有左表数据和左表匹配的右表数据，右表无法匹配的数据用“NULL”补齐。 • right：返回所有右表数据和右表匹配的左表数据，左表无法匹配的数据用“NULL”补齐。 • outer：仅返回左表和右表匹配的数据，不匹配的左表和右表数据全部丢弃。 • inner：对左表和右表进行数据匹配，并返回左表和右表全量数据，左表或右表匹配不上的全部用“NULL”补齐。
左数据列名后缀	左数据中与右数据重复的特征列，加后缀名。支持自定义。
右数据列名后缀	右数据中与左数据重复的特征列，加后缀名。支持自定义。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“数据连接”代码框内容。

----结束

数据联合

数据样本量不足，则无法训练出具有一定泛化能力的模型，训练平台支持具有相同特征维度的数据集进行数据联合，以扩大样本量。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据合并 > 数据联合”。

界面新增如图4-47所示的内容。

图 4-47 数据联合

参数含义如表4-38所示。

表 4-38 数据联合参数说明

参数	参数说明
数据列表	需要进行数据联合的数据，多份数据以逗号分隔。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“数据联合”代码框内容。

----结束

4.6.3.7 特征数据转换

重命名

对特征名称重命名。操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据转换 > 重命名”。

界面新增如图4-48所示的内容。

图 4-48 重命名

参数含义如表4-39所示。

表 4-39 重命名参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： • 列选择 • 正则匹配
列名	单击“…”选定待重命名的特征列，支持至少选择一个列。
正则表达式	列筛选方式为“正则匹配”时展示，请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
新列名	修改后的特征名称。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击图标，运行“重命名”代码框内容。

----结束

归一化

如果一个特征中大部分数据处在（0,100）之间，只有一个数值是10000，或者一个特征的数据分布的区间太长，都有可能会导致模型训练的效果不佳。可通过“归一化”操作将特征值映射到一定的数据区间内，以达到更好的模型训练效果。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据转换 > 归一化”。

界面新增如图4-49所示的内容。

图 4-49 归一化

参数含义如表4-40所示。

表 4-40 归一化参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	列筛选方式为“列选择”时展示，如果多列特征数据均需要归一化到同一数据区间，可单击“…”同时选中多列特征名称。
新列名	默认为空，则直接在原特征列上面做归一化处理。如果设置“新列名”，则原特征列不变，新增经过归一化处理后的一列。
正则表达式	列筛选方式为“正则匹配”时展示，请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
归一化最小值	特征工程归一化后数据均大于“归一化最小值”。 默认值：0。
归一化最大值	特征工程归一化后数据均小于“归一化最大值”。 默认值：1。即特征归一化完成后，数据的区间为(0,1)。
数据最小值	需要做归一化处理的特征数据最小值或者特征理论上可以取到的最小值。如果用户输入，则直接从界面获取，否则后台自动计算特征数据最小值。 默认值为“None”。即用户不输入数据最小值。
数据最大值	需要做归一化处理的特征数据最大值或者特征理论上可以取到的最大值。如果用户输入，则直接从界面获取，否则后台自动计算特征数据最大值。 默认值为“None”。即用户不输入数据最大值。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“归一化”代码框内容。

----结束

数值化

如果特征不是数值型，不利于模型训练。可以通过数值化将其转换为数值型。数值化的思路是根据特征列的样本数据的种类进行编码，数值化后样本数据为取值范围在[0, 样本数据种类-1]区间内的整型数据。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据转换 > 数值化”。

界面新增如图4-50所示的内容。

图 4-50 数值化

参数含义如表4-41所示。

表 4-41 数值化参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	特征名称。单击“…”设置特征列，支持至少选择一列。
正则表达式	列筛选方式为“正则匹配”时展示，请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
新列名	如果设置新列名称，则数值化操作完成后，会生成新特征列，原有特征列保持不变；如果不设置“新列名”，数值化后直接覆盖原有特征列。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击图标，运行“数值化”代码框内容。

----结束

特征离散化

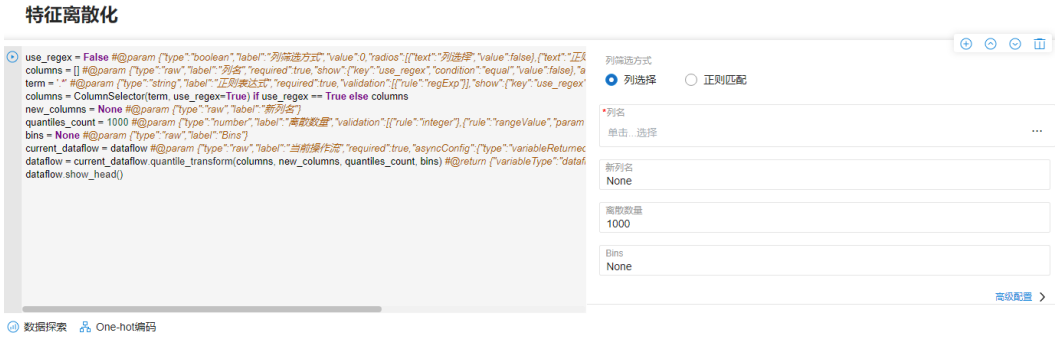
特征离散化是将特征列连续的样本数据离散化为[0，离散数量-1]区间内的整型数据。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据转换 > 特征离散化”。

界面新增如图4-51所示的内容。

图 4-51 特征离散化



参数含义如表4-42所示。

表 4-42 特征离散化参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	特征名称。单击“…”设置特征列，支持至少选择一列。
正则表达式	列筛选方式为“正则匹配”时展示，请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
新列名	如果设置，则特征离散化后生成新特征列，原有特征列不变。如果不设置，则默认覆盖已有特征列。
离散数量	特征数据离散后的取值数量。
Bins	分桶个数。请根据实际情况设置。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“特征离散化”代码框内容。

----结束

One-hot 编码

One-hot编码是根据特征列样本数据的种类对应拆分成相同数量的特征列，将原特征数据映射到新特征中，样本数据相同编码为1，不同编码为0。以特征“Sepal”的样本数据为（2,9,2,8,4）为例，One-hot编码后，会拆分成四列特征，每个特征的样本数据为：

- Sepal_2：10100
- Sepal_4：00001

- Sepal_8：00010
- Sepal_9：01000

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据转换 > One-hot编码”。

界面新增如图4-52所示的内容。

图 4-52 One-hot 编码

参数含义如表4-43所示。

表 4-43 One-hot 编码参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	特征名称。单击“…”设置特征列，支持至少设置一列。
正则表达式	列筛选方式为“正则匹配”时展示，请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
列名前缀	新生成的特征名称前缀。 如果不设置，默认为当前特征名称。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击图标，运行“One-hot编码”代码框内容。

----结束

新增特征

新增特征是对已有特征列进行加、减、乘、除等操作后，生成的新特征。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据转换 > 新增特征”。

界面新增如图4-53所示的内容。

图 4-53 新增特征

参数含义如表4-44所示。

表 4-44 新增特征参数说明

参数	参数说明
表达式	生成新特征的表达式，目前支持对已有特征进行加减乘除、取余、幂方、取模等多种常见运算操作。 支持对多列进行运算生成新特征。
新列名	新特征名称。
在此列前	输入特征名称，则新增的特征展示在此特征之前。 默认值为空，说明新增特征默认放在数据最后一列展示。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“新增特征”代码框内容。

----结束

Box-Cox 变换

用于连续的响应变量不满足正态分布时，进行数据变换，达到接近正态分布的目的。Box-Cox变换的主要特点是引入一个参数，通过数据本身估计该参数，进而确定应采取的数据变换形式。

使用Box-Cox变换的优点：

- 数据得到的回归模型优于变换前的模型，变换可以使模型的解释力度等性能更加优良。
- 降低偏度值，残差可以更好的满足正态性、独立性等假设前提，使其更加符合后续对数据分布的假设，降低了伪回归的概率。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 数据转换 > Box-Cox变换”。

界面新增如图4-54所示内容。

图 4-54 Box-Cox 变换

参数含义如表4-45所示。

表 4-45 Box-Cox 变换参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	特征名称。单击“…”设置特征列，支持至少设置一列。
正则表达式	列筛选方式为“正则匹配”时展示，请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
变换参数	Box-Cox变换的变换参数值，需为数字，默认为空。如果为空，则自动寻找最优变换参数值；如果为数字，则“列名”选择的所有特征列均使用此值。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“Box-Cox变换”代码框内容。

----结束

4.6.3.8 特征数据选择

删除列

删除特征列的场景有很多，例如：两个特征呈线性变化关系，为减少模型训练的开销，删除其中一个特征列。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 特征选择 > 删除列”。

界面新增如图4-55所示的内容。

图 4-55 删除列

参数含义如表4-46所示。

表 4-46 删除列参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	列筛选方式为“列选择”时展示，如果有多列特征数据需要删除，可单击“…”同时选中多列特征名称。
正则表达式	列筛选方式为“正则匹配”时展示，请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击图标，运行“删除列”代码框内容。

----结束

选择列

如果数据的特征量大，而大多数特征对模型训练无效，可通过“选择列”保留仅对模型训练有意义的特征。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 特征选择 > 选择列”。

界面新增如图4-56所示的内容。

图 4-56 选择列

参数含义如表4-47所示。

表 4-47 选择特征参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	列筛选方式为“列选择”时展示，如果有多列特征数据需要保留，可单击“…”同时选中多列特征名称。
正则表达式	列筛选方式为“正则匹配”时展示，请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“选择列”代码框内容。

----结束

4.6.3.9 时序特征数据处理

缺失时间填充

时序序列是在连续的等间隔时间点采集的序列，缺失时间填充即根据已知的时间信息，补充缺失的时间。缺失时间填充完成后，其值可通过“数据处理 > 数据清洗 > 空值填充”菜单，进行空值填充。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 时序数据处理 > 缺失时间填充”。

界面新增如图4-57所示的内容。

图 4-57 缺失时间填充

参数含义如表4-48所示。

表 4-48 缺失时间填充参数说明

参数	参数说明
时间列	待填充缺失时间特征数据的时间列。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。

参数	参数说明
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击图标，运行“缺失时间填充”代码框内容。

----结束

时序数据排序

时序数据排序即根据给定的参数对时间序列进行排序。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 时序数据处理 > 时序数据排序”。

界面新增如图4-58所示的内容。

图 4-58 时序数据排序

参数含义如表4-49所示。

表 4-49 时序数据排序参数说明

参数	参数说明
时间列	时序数据时间列，系统将根据指定的时间，按时间从早到晚对时序数据进行排序。
ID列	时序数据的标识列，默认为空，如果指定ID列，系统将按照（ID，Time）的方式对时序数据进行升序排序。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击图标，运行“时序数据排序”代码框内容。

----结束

时间迁移

时间迁移即转换时序数据的时间，如将时间整体向前推移或整体向后推移等。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 时序数据处理 > 时间迁移”。

界面新增如图4-59所示的内容。

图 4-59 时间迁移

参数含义如图4-59所示。

表 4-50 时间迁移参数说明

参数	参数说明
时间列	待迁移时间的时间字段。
迁移量	迁移的幅度，如“-3min9s”表示指定时间列值减去3分9秒；“2h30min”表示指定时间列值加2小时30分。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“时间迁移”代码框内容。

----结束

时序数据重采样

时序数据重采样即时间序列从一个频率转换到另一个频率的过程。

其中：

- 高频率（采样间隔短）数据转换到低频率（采样间隔长）称为降采样。
- 低频率数据转换到高频率称为升采样。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 时序数据处理 > 时序数据重采样”。

界面新增如图4-60所示的内容。

图 4-60 时序数据重采样

参数含义如表4-51所示。

表 4-51 时序数据重采样参数说明

参数	参数说明
时间列	时序数据的时间字段。
重采样频率	重采样时间频率，如“5H”。 时间频率单位说明： ● S：秒 ● min：分钟 ● H：小时 ● D：天 ● B：工作日 ● W：周 ● M：月 ● Q：季 ● A：年
重采样方法	当前支持的重采样方法： ● 升采样时可选择：不填充、前向填充、后向填充、插值填充。 ● 降采样时可选择：求和、求均值、求方差、中位数、第一个值、最大值、最小值、最后一个值。 如果采样方法为空，则升采样默认方法为不填充；降采样默认方法为均值聚合。采样方法支持传入自定义函数。
ID列	时序数据的标识列。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“时序数据重采样”代码框内容。

----结束

时序数据去噪

时序数据中可能会存在许多噪声数据，这些噪声严重影响进一步的定量分析和数据挖掘，因此需要进行数据去噪。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 时序数据处理 > 时序数据去噪”。界面新增如图4-61所示的内容。

图 4-61 时序数据去噪

参数含义如表4-52所示。

表 4-52 时序数据去噪参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	列筛选方式设置为“列选择”时才会展示。 通过单击“…”图标，在弹出的对话框中，选择一个或者多个特征列。
正则表达式	列筛选方式设置为“正则匹配”时才会展示。 请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
新列名	经过去噪后产生的新数据的列名。如果不设置，则直接在原有特征列上进行去噪处理。
时间列	待去噪时序数据的时间列。
其他参数配置	该参数用于在去噪时指定frac值。 去噪使用了statsmodels的局部加权回归散点平滑法（locally weighted scatterplot smoothing, LOWESS），其中局部表示每次只处理数据的一部分，此部分数据所占整体的比例由LOWESS的frac参数表示，而frac值可通过该参数传递。具体用法可参见查看“帮助中心 > SDK文档”。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击图标，运行“时序数据去噪”代码框内容。

----结束

时间特征提取

时间特征提取是指从时序数据的时间列中提取出日期相关的特征，如年、月、日、时、分、秒、季节、星期几、一年中的第几周、一年中的第几天等特征。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 时序数据处理 > 时间特征提取”。

界面新增如图4-62所示的内容。

图 4-62 时间特征提取

参数含义如表4-53所示。

表 4-53 时间特征提取参数说明

参数	参数说明
时间列	待进行时间特征提取的时间列。
预提取时间特征	要提取的时间特征。默认为“全量提取”，指提取全部的时间特征。此外还支持提取“年”、“月”、“日”、“时”、“分”、“秒”、“星期几”、“一年中的第几天”、“一年中的第几周”、“季”这些时间特征。
新列名	提取出时间特征后产生的新特征列的列名。如果不设置，则默认采用时间列名称加特征名称的命名方式。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击 图标，运行“时间特征提取”代码框内容。

----结束

时序数据特征提取

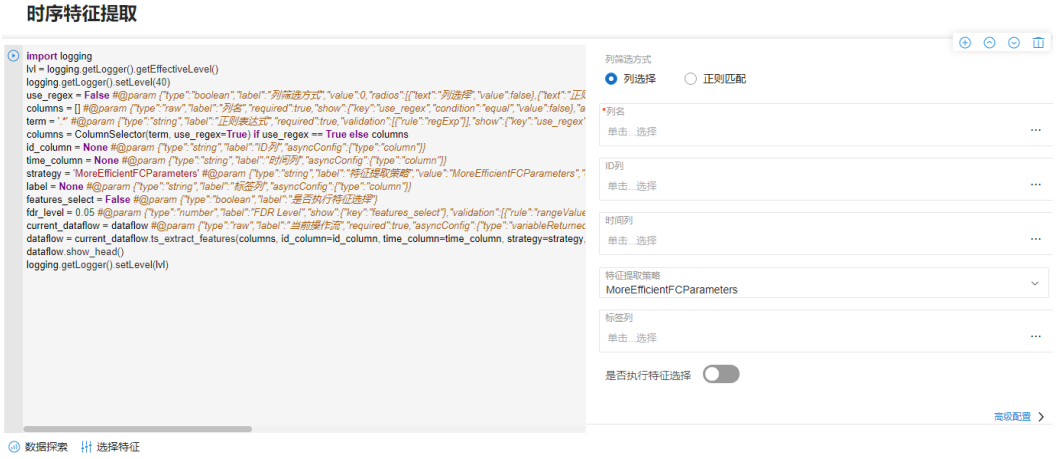
时序特征提取，即从时序数据中提取数据统计学特性，最大限度地找出样本内时间序列的统计特性和发展规律。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 时序数据处理 > 时序特征提取”。

界面新增如图4-63所示的内容。

图 4-63 时序特征提取



参数含义如表4-54所示。

表 4-54 时序特征提取参数说明

参数	参数说明
列筛选方式	特征列的筛选方式，有如下两种： ● 列选择 ● 正则匹配
列名	列筛选方式为“列选择”时展示，时序特征提取的目标特征列，可单击“…”从特征列中选择一个或多个特征列。
正则表达式	列筛选方式为“正则匹配”时展示，请根据实际情况输入正则表达式，系统自动筛选符合正则筛选规则的所有特征列。
ID列	单击“…”从特征列中选取特征标识字段作为时序特征提取的ID列，仅支持单列选取。系统会根据ID列进行分组特征提取，如果不设置ID列，则默认“列名”选择的所有列数据都具有相同的ID。
时间列	单击“…”从特征列中选取时间字段作为时序特征提取的时间列，仅支持单列选取。如果为空，则认为时序数据已经按时间顺序排列。
特征提取策略	特征提取分层参数配置策略，支持如下策略： ● SmallEfficientFCParameters ● MoreEfficientFCParameters ● CombinedFCParameters
是否执行特征选择	是否选择提取的特征。
标签列	单击“…”从特征列中选取一列作为标签列，指定用于分析其他特征列和标签列的相关性。

参数	参数说明
FDR Level	“是否执行特征选择”开启时展示，进行特征选择时使用，表示显著性水平，是理论上的预期不相关特征在所有特征中所占的百分比。默认值为“0.05”。
当前操作流	高级配置参数，从下拉框中选择当前输入操作流的名字。
重命名操作流	高级配置参数，重命名当前输出操作流的名字。

步骤2 单击图标，运行“时序特征提取”代码框内容。

----结束

4.6.3.10 自定义特征处理

自定义操作

提供特征处理代码编辑能力，满足用户自定义特征处理需求。

操作步骤如下所示。

步骤1 在特征工程“算子”菜单栏中，选择“数据处理 > 自定义 > 自定义操作”。

界面新增如图4-64所示的内容。

图 4-64 自定义特征操作

步骤2 在“Your code here”注释行下方输入自定义特征操作的代码。

如需重命名操作流输出变量名称，可展开“高级配置”，修改“dataflow”参数值，默认值为“dataflow”。

步骤3 单击图标，运行“自定义操作”代码框内容。

----结束

自定义算子

支持用户进行自定义开发，新增算法文件，然后在后缀名称为“.ipynb”文件中调用算法并执行。

也可以直接在“自定义算子”模块设置运行。

自定义算子开放训练和推理侧定制能力，支持维护上下文，自定义算子推理侧可复用。自定义算子的代码要求，可从SDK文档的“特征处理 > 自定义算子”章节查看。

4.6.3.11 全量数据应用

特征操作完成后，需要选择“算子 > 数据处理 > 数据集 > 生成数据实例”，应用特征操作流至全量数据，并生成经过特征处理后的新数据。详细请参见**生成数据实例**。

4.6.3.12 发布服务

如果当前特征工程操作流处理效果比较好，可以得到比较优质的训练数据，可以将当前的特征工程发布成服务。复用此特征工程服务对其他数据进行相同的特征操作。

步骤1 在特征工程菜单栏中，单击 发布。

步骤2 在弹出的“Publish”框内设置发布后的“Service Name”。

步骤3 单击“Publish”。

步骤4 在弹出的“Success”框内单击“OK”。

发布后的工程可在特征工程首页的“已发布服务”页签内查看。

发布后的特征工程可创建特征工程任务，详细操作请参见“**创建特征工程任务**”。

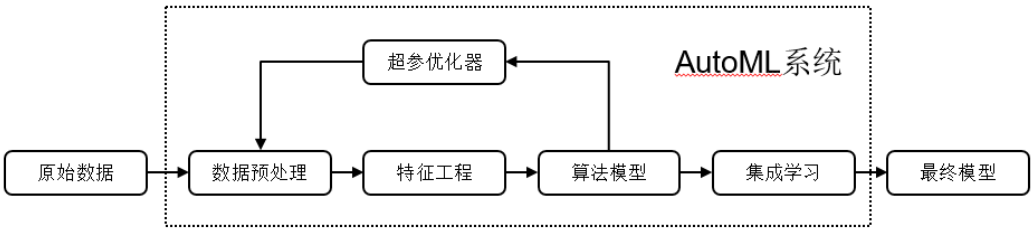
----结束

4.6.3.13 基于 Jupyterlab 的自动机器学习

AutoML

AutoML（VegaAutoML）是以华为诺亚实验室VegaAutoML为原型开发的SDK，便于开发者使用和开发AI模型。AutoML采用业界经典的AutoML框架，主要包括数据预处理、特征工程、算法模型、超参优化、集成学习五个模块。其中，超参优化模块是对数据预处理、特征工程、算法模型构成的pipeline进行超参寻优。AutoML框架图如**图4-65**所示。

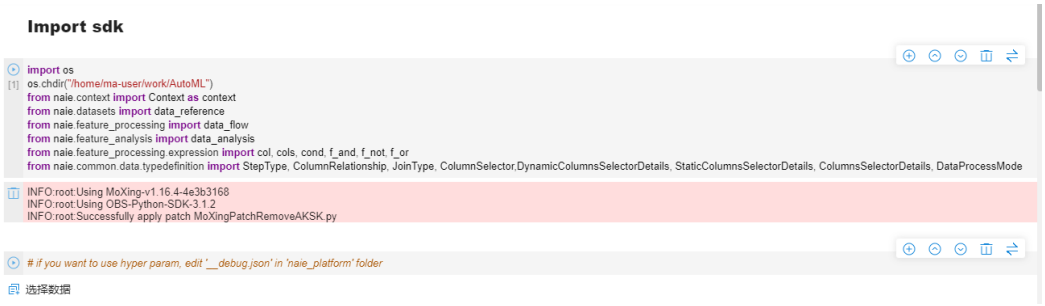
图 4-65 AutoML 框架图



下面以系统预置的样例数据为例，进行AutoML操作演示。

步骤1 单击如所示的 图标，运行“Import sdk”内容。

图 4-66 导入 SDK



步骤2 单击“Import sdk”右侧的 图标，新增cell。

输入如下代码：

```
from naie.datasets import samples
samples.load_dataset("higgs", "higgs_train_10k")
samples.load_dataset("higgs", "higgs_test_5k")
```

步骤3 单击新增cell左侧的 图标，加载两份higgs数据集分别作为训练集和测试集，如[图4-67](#)所示。

图 4-67 加载训练集

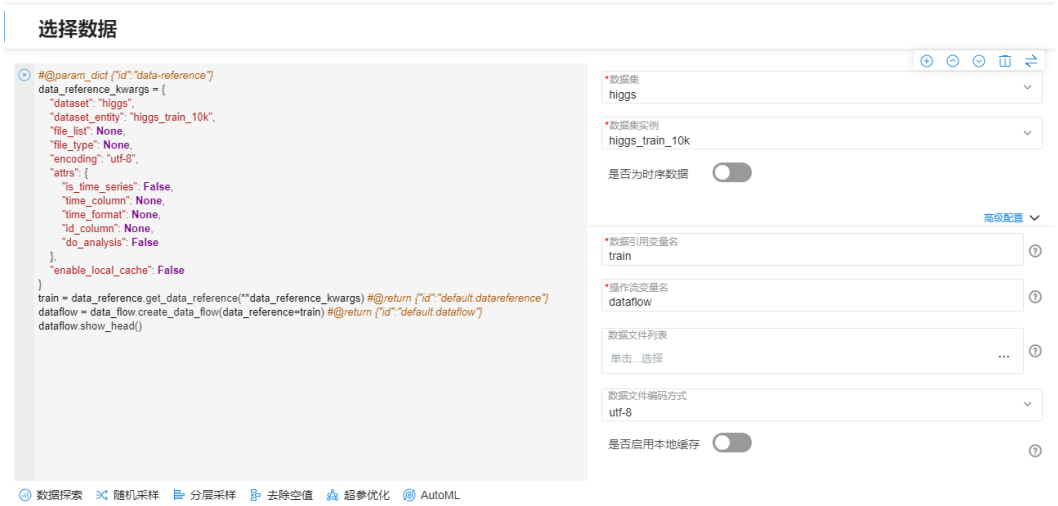
步骤4 在界面右上角，选择“算子 > 数据处理 > 数据集 > 选择数据”。

新增“选择数据”内容，如[图4-68](#)所示。

设置如下参数取值，其余参数保持默认值即可。

- 数据集：从下拉框中选择“higgs”。
- 数据集实例：从下拉框中选择“higgs_train_10k”。
- 数据引用变量名：方便后面通过此变量名称，引用当前数据，示例为“train”。

图 4-68 选择数据



步骤5 单击 图标，运行“选择数据”代码框内容。训练集绑定成功。

步骤6 请参考**步骤4**和**步骤5**操作，绑定测试集。

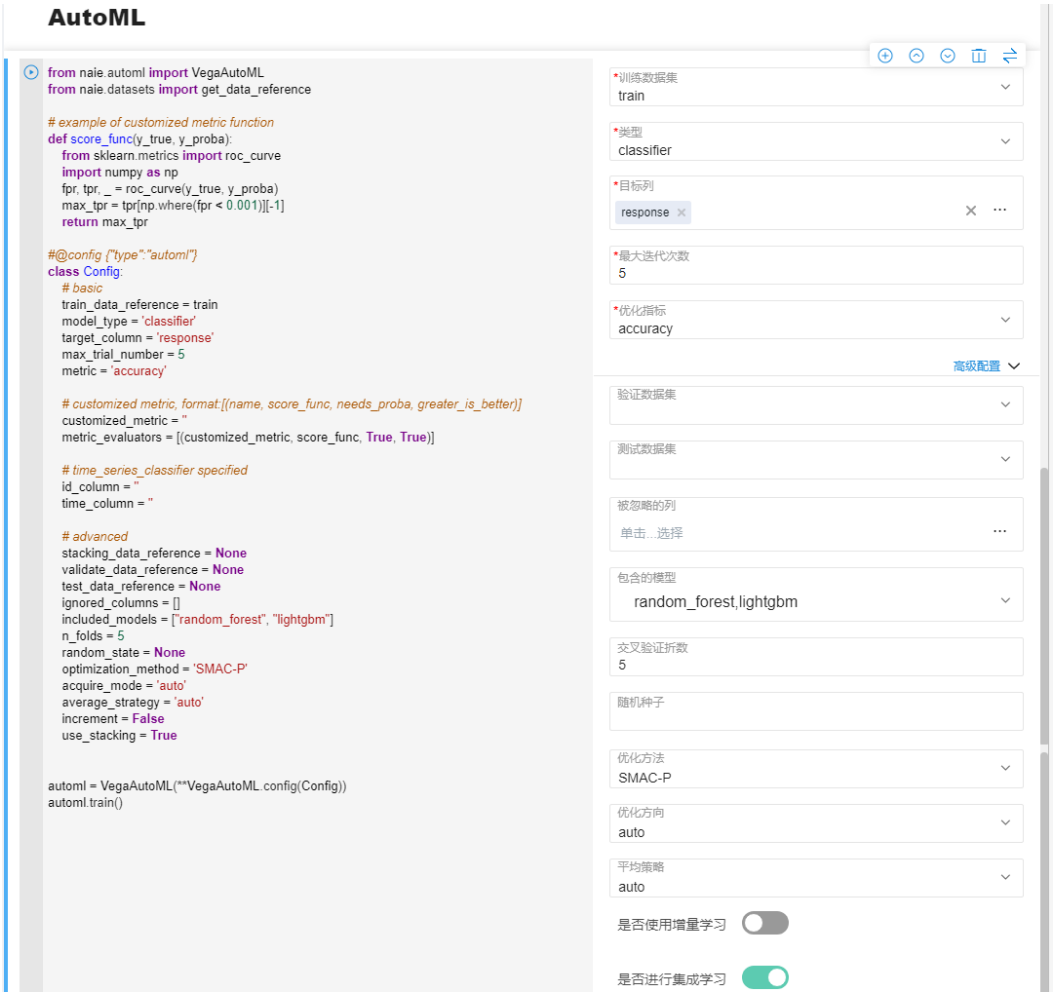
下述参数，对应修改为：

- 数据集实例：选择“higgs_test_5k”。
- 数据引用变量名：设置为“test”。

步骤7 在界面右上角，选择“算子 > 模型训练 > 模型训练 > AutoML”。

界面新增如**图4-69**所示的内容。

图 4-69 AutoML 参数设置



参数设置说明，如表4-55所示。

表 4-55 AutoML 参数说明

参数	参数说明
训练数据集	训练数据集。从下拉框中选择“train”，即步骤4中的“数据引用变量名”。
类型	训练的模型类型。 目前支持如下类型： • classifier：分类类型 • regressor：回归类型 • time_series_classifier：时序分类类型；如果选择此类型，默认新增两个配置参数“标识列”和“时间列”。其中“标识列”的含义为：标记哪些数据是属于同一对象的，为必填参数；“时间列”的含义为：同一个对象的数据排序。 当前样例数据用于生成分类类型的模型，请选择“classifier”。

参数	参数说明
目标列	数据的标签列。必填参数。设置为“reponse”。
最大迭代次数	AutoML任务中模型训练迭代次数上限。默认值为“5”。
优化指标	AutoML任务的模型优化指标，请根据实际情况选择。
验证数据集	模型验证数据集。
测试数据集	模型测试数据集。
被忽略的列	数据集中不需要参与模型训练的无用列。
包含的模型	模型训练使用的算法列表。
交叉验证折数	交叉检验的折数。如果不使用交叉验证方法，请将该参数置为空。 K折交叉验证的含义：将数据集等比例划分成K份，其中一份作为测试数据，其他的（K-1）份数据作为训练数据，这样算是一次实验。K折交叉验证为实验K次才算完成的一次，即保证K份数据分别做过测试数据。最后把得到的K个实验结果进行评分。 保持默认值即可。
随机种子	以一个真随机数（随机种子）作为初始条件，使用一定的算法不停迭代产生随机数。
优化方法	超参优化方法。 目前支持如下方法： ● GPEI ● GPTS ● SMAC ● SMAC-P
优化方向	超参优化的目标。 包含如下选项： ● auto ● max ● min 默认值：auto。

参数	参数说明
平均策略	计算指标的平均策略。 包含如下选项： • auto • macro • micro • weighted
是否使用增量学习	训练时是否使用增量学习，默认关闭。
是否进行集成学习	训练时是否进行集成学习，默认开启。开启后训练结果增加模型集成节点，训练结果中生成两个stacking类型的模型包。

步骤8 单击图标，运行AutoML代码框内容。运行结果如图4-70所示。

AutoML模型训练过程中，会展示“AutoML过程设置”内容，用户可调整此区域的参数设置，重新选择使用的模型，或关闭特征搜索。

其中“排行榜”展示所有训练出的模型列表，支持对模型进行如下操作：

- 单击模型所在行对应“操作”列的“详情”，查看模型超参取值和模型评分结果。
- 单击模型所在行对应“操作”列的“预测”，在新增的“AutoML模型预测”内容中，选择测试数据集test，运行代码框，查看模型预测结果，如图4-71所示。

  通常使用最优模型，配合测试数据集进行结果预测和模型评分，查看测试结果是否符合预期。
- 单击模型所在行对应“操作”列的“保存”，保存当前模型。可在左侧文件目录“特征处理工程名称/debug/output”目录下查看同名模型包文件。

图 4-70 AutoML 运行结果

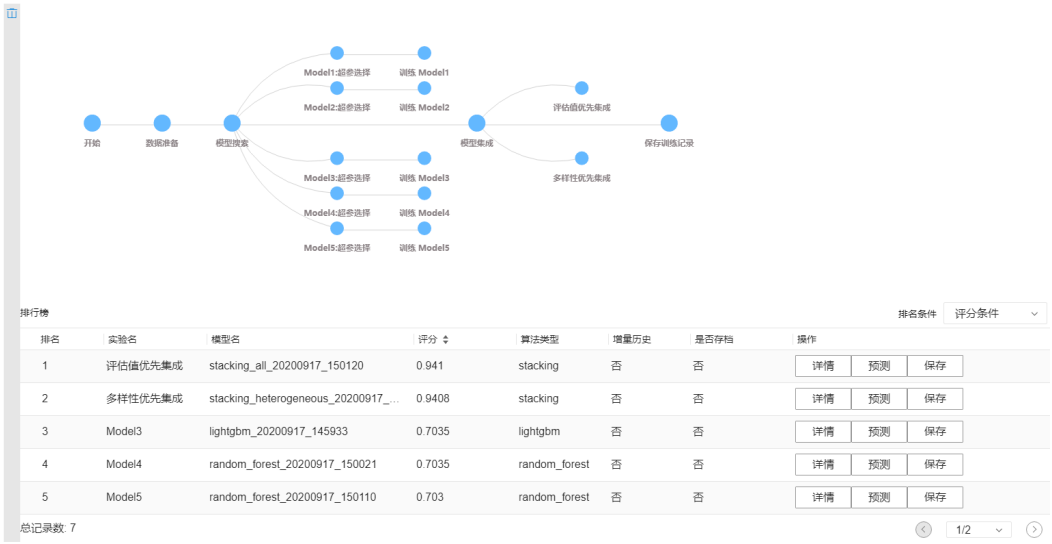
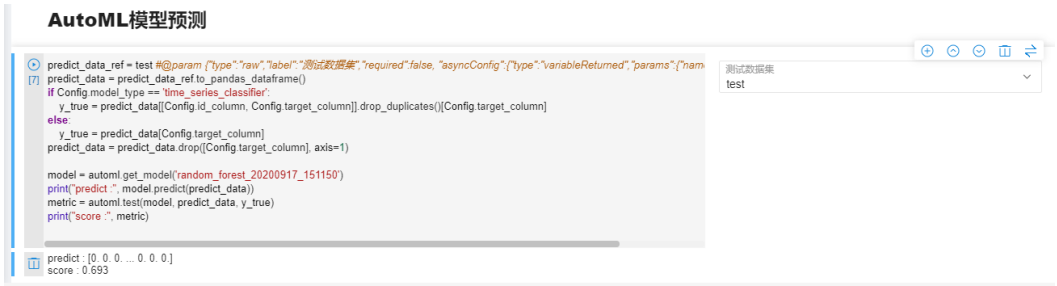


图 4-71 模型预测结果



----结束

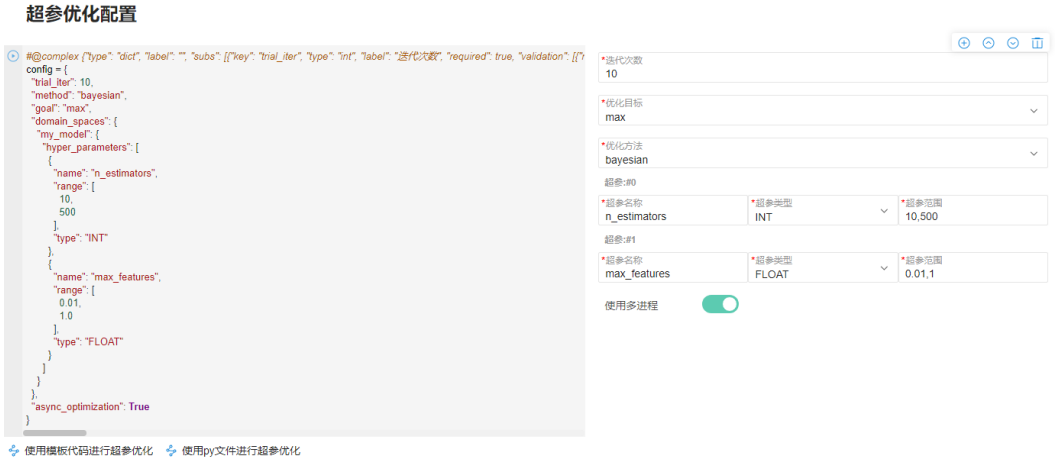
超参优化

超参优化是对数据预处理、特征工程、算法模型构成的pipeline进行超参寻优。这一过程不仅需要专家经验，还会耗费大量时间。使用超参优化功能可以快速、自动、高效地找到最优模型超参，帮助用户节约时间，降低工作复杂度。

步骤1 在特征工程“算子”菜单栏中，选择“模型训练 > 模型训练 > 超参优化”。

界面新增“超参优化配置”代码框。“超参优化配置”如图4-72所示。

图 4-72 超参优化配置



超参优化配置参数含义如表4-56所示。

表 4-56 超参优化配置参数说明

参数	参数说明
迭代次数	超参优化任务的最小迭代次数。
优化目标	超参优化任务的目标，在训练算法中进行定义，支持“max”和“min”两个目标。

参数	参数说明
优化方法	超参优化方法： • smac • bayesian • random • grid
超参名称	超参名称，可根据算法自定义设置。
超参类型	超参的类型，请根据实际情况选择超参类型。
超参范围	超参的取值区间，请根据实际需要设置超参最小值和最大值。
使用多进程	超参优化过程是否启动多进程，默认开启。

步骤2 单击“超参优化配置”对应的图标，运行代码框内容。

步骤3 使用模板代码进行超参优化。

1. 单击“超参优化配置”代码框下方的“使用模板代码进行超参优化”，弹出如图4-73所示内容。

   图 4-73 使用模板代码进行超参优化

2. 从特征列中选取标签列，然后单击“使用模板代码进行超参优化”对应的图标，运行代码框内容。

步骤4 使用py文件进行超参优化。

1. 右键单击特征工程左侧目录列表空白区域，选择“New File”新增一个名为“train.py”的主函数文件，并在该文件中定义主函数。

   主函数内容参考如下：

```
from naie.context import Context
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split

def train_func():
    iris = load_iris()
    X_train, X_test, y_train, y_test = train_test_split(iris.data, iris.target)
    model = RandomForestClassifier(n_estimators = Context.get("n_estimators"), max_features =
Context.get("max_features"))
    model.fit(X_train, y_train)
    y_pred = model.predict(X_train)
    return accuracy_score(y_train, y_pred)
```


2. 单击“超参优化配置”代码框下方的“使用py文件进行超参优化”，弹出如图4-74所示内容。

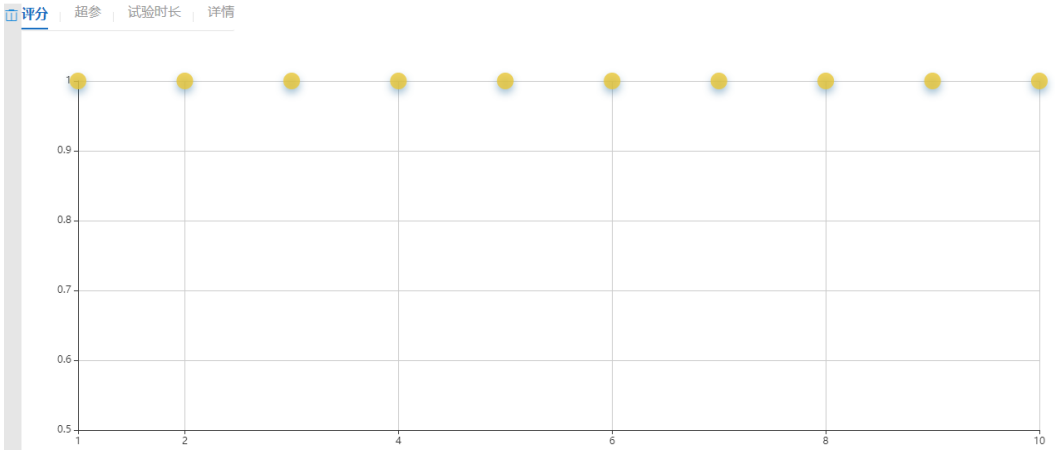
   “文件”和“函数”自动回填为已定义的主函数文件名及主函数文件内定义的主函数名。

图 4-74 使用 py 文件进行超参优化

3. 单击“使用py文件进行超参优化”对应的图标，运行代码框内容。

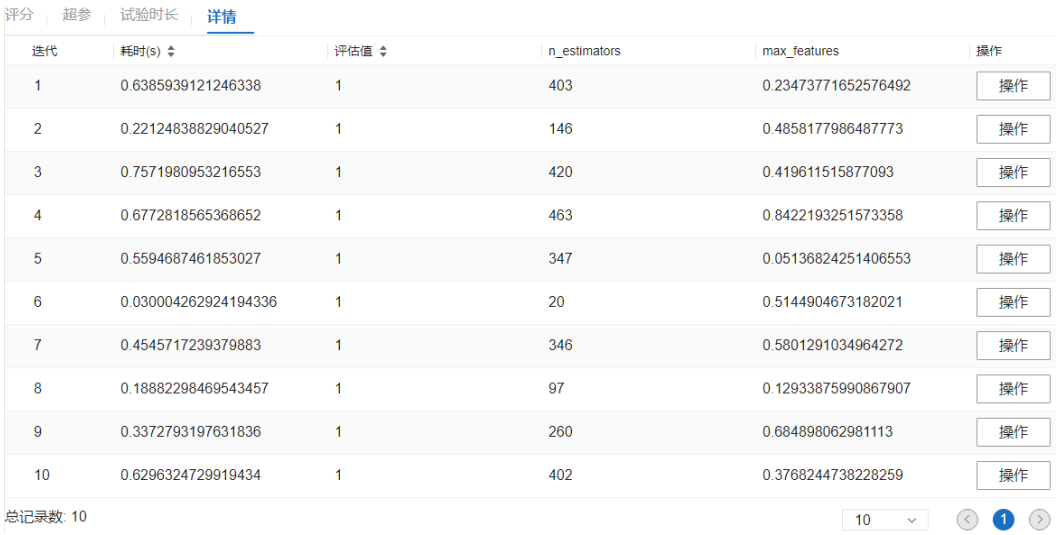
   运行成功后，可查看“评分”、“超参”、“试验时长”以及“详情”四个超参优化结果，如图4-75所示。

图 4-75 使用 py 文件的超参优化结果



4. 单击图4-75的“详情”页签，该页签展示模型评分、训练耗时、超参优化参数及其取值信息，如图4-76所示。

图 4-76 使用 py 文件的超参优化结果详情



5. 单击其中一个模型操作列对应的“操作”，提取超参做进一步操作，如图4-77所示。

图 4-77 超参优化模型操作

6. 从特征列中选取标签列，然后单击“超参优化模型操作”对应的图标，运行代码框内容。

----结束

4.6.3.14 特征迁移

如果当前数据集的特征数据不够理想，而此数据集的数据类别和一份理想的数据集部分重合或者相差不大的时候，可以使用特征迁移功能，将理想数据集的特征数据迁移到当前数据集中。

进行特征迁移前，请先完成如下操作：

- 将源数据集和目标数据集导入系统，详细操作请参见**数据集**。
- 创建迁移数据Jupyterlab特征工程，详细操作请参见**创建特征工程**。

注意

请按照本章章节的操作顺序在特征工程中完成数据迁移，若其中穿插了其他数据操作，需要保证有前后衔接关系的两个代码框的dataflow名字一致。

绑定源数据

步骤1 进入迁移数据特征工程编辑界面，运行“Import sdk”代码框。

步骤2 在特征工程“算子”菜单栏中，选择“迁移学习 > 特征迁移 > 特征准备 > 绑定源数据”。

界面新增如图4-78所示的内容。

图 4-78 绑定迁移前的源数据

参数含义如表4-57所示。

表 4-57 绑定迁移前的源数据参数说明

参数	参数说明
数据集	迁移前源数据对应的数据集。
数据集实例	迁移前源数据的数据集实例。
源数据引用变量名	修改源数据引用变量名，以免和目标数据引用变量名产生冲突。当有多份数据需要迁移时，也可作为同类数据之间引用变量名的区分。
源操作流变量名	修改源操作流变量名，以免和目标操作流变量名产生冲突。当有多份数据需要迁移时，也可作为同类数据之间操作流变量名之间的区分。

步骤3 单击 图标，运行“绑定迁移前的源数据”代码框内容。

----结束

绑定目标数据

步骤1 在特征工程“算子”菜单栏中，选择“迁移学习 > 特征迁移 > 特征准备 > 绑定目标数据”。

界面新增如图4-79所示的内容。

图 4-79 绑定迁移前的目标数据

参数含义如表4-58所示。

表 4-58 绑定迁移前的目标数据参数说明

参数	参数说明
数据集	迁移前目标数据对应的数据集。
数据集实例	迁移前目标数据的数据集实例。
目标据引用变量名	修改目标数据引用变量名，以免和源数据引用变量名产生冲突。当有多份数据需要迁移时，也可作为同类数据之间引用变量名的区分。

参数	参数说明
目标操作流变量名	修改目标操作流变量名，以免和源操作流变量名产生冲突。当有多份数据需要迁移时，也可作为同类数据之间操作流变量名之间的区分。

步骤2 单击图标，运行“绑定迁移前的目标数据”代码框内容。

----结束

评估迁移数据

在使用迁移算法对数据进行迁移前，可以使用评估迁移数据功能评估当前数据是否适合迁移。

步骤1 在特征工程“算子”菜单栏中，选择“迁移学习 > 特征迁移 > 迁移评估 > 评估迁移数据”。

界面新增如图4-80所示内容。

图 4-80 评估迁移数据

参数含义如表4-59所示。

表 4-59 评估迁移数据参数说明

参数	参数说明
源操作流变量名	对应绑定迁移前源数据设置的源操作流变量名。
目标操作流变量名	对应绑定迁移前目标数据设置的目标操作流变量名。

步骤2 根据实际源数据集和目标数据集标签列的值修改左侧代码区域中“# Select data from dataframe”下SX和TX的值。

步骤3 单击图标，运行“评估迁移数据”代码框内容。

----结束

评估迁移算法

如果评估迁移数据的结果为当前数据适合迁移，可以使用评估迁移算法评估当前数据适合采用哪种算法进行迁移。

步骤1 在特征工程“算子”菜单栏中，选择“迁移学习 > 特征迁移 > 迁移评估 > 评估迁移算法”。

界面新增如图4-81所示内容。

图 4-81 评估迁移算法

参数含义如表4-60所示。

表 4-60 评估迁算法据参数说明

参数	参数说明
源操作流变量名	对应绑定迁移前源数据设置的源操作流变量名。
目标操作流变量名	对应绑定迁移前目标数据设置的目标操作流变量名。

步骤2 根据实际源数据集和目标数据集标签列的值修改左侧代码区域中“# Select data from dataframe”下SX、SY和TX的值。

步骤3 单击图标，运行“评估迁移算法”代码框内容。

----结束

迁移操作

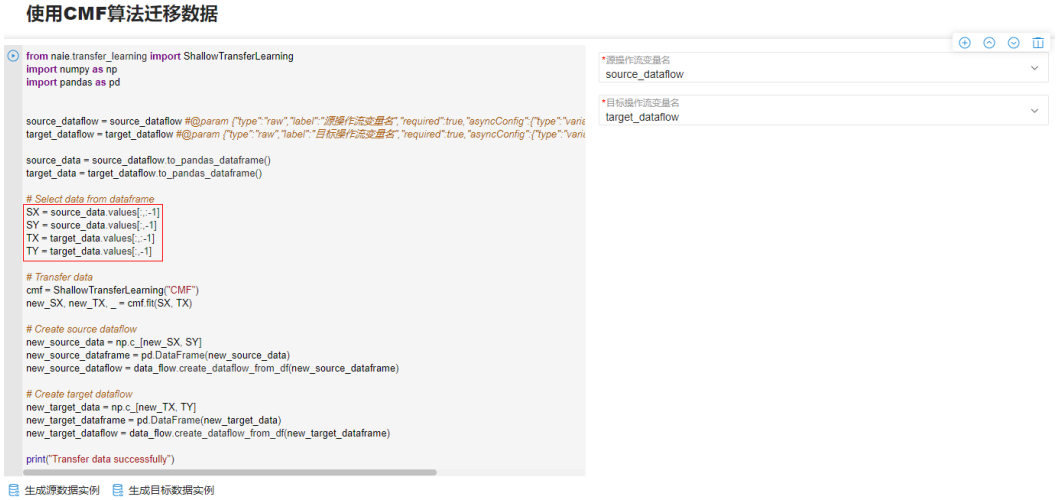
当前系统支持的迁移算法有：CMF、CORAL、GFK、ITL、KMM、LSDT、MSDA、PCA、RANDPROJ、SA、TCA，每种算法不需要另外设置参数，只需用户根据实际源数据和目标数据的标签列修改代码框左侧“# Select data from dataframe”标注下的对应值。

本文以使用“CMF”方法为例。

步骤1 在特征工程“算子”菜单栏中，选择“迁移学习 > 特征迁移 > 迁移操作 > CMF”。

界面新增如图4-82所示内容。

图 4-82 使用 CMF 算法迁移数据



参数含义如表4-61所示。

表 4-61 使用 CMF 算法迁移数据参数说明

参数	参数说明
源操作流变量名	对应绑定迁移前源数据设置的源操作流变量名。
目标操作流变量名	对应绑定迁移前目标数据设置的目标操作流变量名。

步骤2 根据实际源数据集和目标数据集标签列的值修改图4-82红框区域对应值。其中，S表示源数据，T表示目标数据，X表示数据特征，Y表示数据标签。

步骤3 单击 图标，运行“使用CMF算法迁移数据”代码框内容。

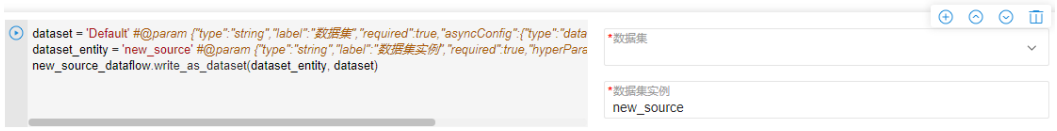
----结束

生成源数据实例

步骤1 在特征工程“算子”菜单栏中，选择“迁移学习 > 特征迁移 > 生成数据 > 生成源数据实例”。

界面新增如图4-83所示内容。

图 4-83 生成迁移后的源数据实例



参数含义如表4-62所示。

表 4-62 生成迁移后的源数据实例参数说明

参数	参数说明
数据集	迁移后源数据对应的数据集。
数据集实例	源数据迁移后生成的数据集实例名，可自定义命名。

步骤2 单击图标，运行“生成迁移后的源数据实例”代码框内容。

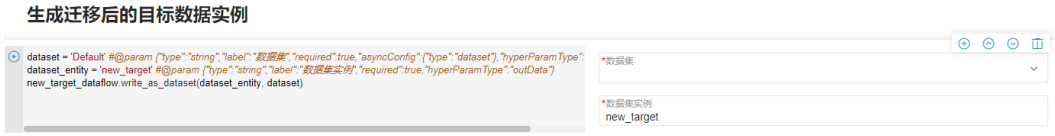
----结束

生成目标数据实例

步骤1 在特征工程“算子”菜单栏中，选择“迁移学习 > 特征迁移 > 生成数据 > 生成目标数据实例”。

界面新增如图4-84所示内容。

图 4-84 生成迁移后的源数据实例



参数含义如表4-63所示。

表 4-63 生成迁移后的源数据实例参数说明

参数	参数说明
数据集	迁移后目标数据对应的数据集。
数据集实例	目标数据迁移后生成的数据集实例名，可自定义命名。

步骤2 单击图标，运行“生成迁移后的目标数据实例”代码框内容。

----结束

4.6.3.15 学件

多层嵌套异常检测学件和硬盘故障根因分析学件的操作，请参考《学件开发指南》。

4.7 模型训练

4.7.1 模型训练简介

训练平台支持所有主流算法框架，如：Tensorflow， MXNet ，Caffe，Spark_MLlib，Scikit_Learn，XGBoost，PyTorch、Ascend-Powered-Engine等。提供CPU、GPU等多种计算资源，集成了基于开源的交互式开发调试工具，为用户提供一站式IDE模型训练环境。

模型训练提供如下功能：

- 新建模型训练工程：支持用户在线编辑并调试代码，基于编译成功的代码对模型训练工程的数据集进行训练，输出训练报告。用户可以根据训练报告结果对代码进行调优再训练，直到得到最优的训练代码。
- 新建联邦学习工程：支持用户在线编辑并调试代码，基于编译成功的代码对联邦学习工程的数据集进行训练，输出训练报告。用户可以根据训练报告结果对代码进行调优再训练，直到得到最优的训练代码。
- 新建训练服务：调用已归档的模型包，对新的数据集进行训练，得到训练结果。
- 新建超参优化服务：通过训练结果对比，为已创建的训练工程选择一组最优超参组合。

系统还支持打包训练模型，用于创建训练服务、模型验证，或者发布到应用市场。模型训练包包括编排配置文件、模型文件等。详细的模型管理操作请参见**模型管理**。

模型训练页面说明

“模型训练”页面列出了已有的训练工程、训练服务和超参优化服务的列表信息，如**图4-85**所示。在该页面，用户可以查看训练工程和训练服务的创建信息，新建、编辑、复制或删除已创建的训练工程和训练服务。详情请参见**表4-64**。

图 4-85 模型训练

表 4-64 模型训练页面说明

参数名称	参数说明
开发环境	WEB IDE环境资源配置，包括配置“规格”和“实例”，用于模型开发。支持查看当前所有配置了WEB IDE环境资源的项目的环境信息。
创建	新建训练工程、联邦学习工程、训练服务或超参优化服务。
名称	模型训练名称。
模型训练工程描述	对模型训练工程的描述信息。
创建时间	训练工程、联邦学习工程、训练服务或者超参优化服务的创建时间。

参数名称	参数说明
类型	模型训练的类型。 包含如下选项： • 模型训练 • 联邦学习 • 训练服务 • 优化服务
创建者	创建训练工程、联邦学习工程、训练服务或超参优化服务的用户。
开发环境	模型训练运行的环境信息。WEB版训练模型的开发环境为“简易编辑器”，在线IDE版训练模型的开发环境为实际创建的WEB IDE环境。模型训练工程创建后，可通过“开发环境”下拉框切换环境。
	进入训练工程编辑页面，编辑训练代码。
	复制已有的训练工程，生成新的训练工程。
	删除训练工程、联邦学习工程、训练服务或优化服务。
FINISHED	最近一次训练工程、联邦学习工程、训练服务或超参优化服务的任务状态。显示实际任务状态。

4.7.2 创建模型训练工程

4.7.2.1 创建工程

创建训练工程是从创建模型训练工程、编辑模型训练代码到调试模型训练代码的端到端的代码开发过程。

- 创建模型训练工程：创建模型训练代码编辑和调试的环境。
- 编辑模型训练代码：在线编辑模型训练代码。
- 调试模型训练代码：在线调试编辑好的模型训练代码。

创建训练工程步骤如下。

步骤1 单击“创建”，弹出“创建训练”对话框。

步骤2 配置训练工程参数，如表4-65所示。

表 4-65 新建训练工程参数说明

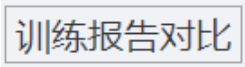
参数名称	参数说明
请选择模型训练方式	模型训练方式。包含如下选项： ● 新建模型训练工程 ● 新建联邦学习工程 ● 新建训练服务 ● 新建超参优化服务 请选择：新建模型训练工程。
模型训练名称	模型训练名称。 只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线（_）组成，不能以下划线结尾，长度范围为[1,26]。
描述	对新建模型训练工程的描述。
模型试验算法	通用算法选择：分类算法、拟合算法、聚类算法、其他类型。如果选择分类算法，可以看到“创建入门模型训练代码”，如果勾选，则自动生成鸢尾花分类建模的样例代码。
开发环境	训练工程使用的开发环境，支持： ● WebIDE WebIDE提供类似本地VSCode的编码体验，支持代码自动补齐、调试等功能，适用于大量代码编写场景。创建在线IDE版训练模型时选择“WebIDE”开发环境。 ● 简易编辑器 简易编辑器提供代码查看和编辑能力，不支持调试，适用于少量代码修改场景。创建WEB版训练模型时，选择“简易编辑器”开发环境。
规格	当“开发环境”选择“WebIDE”时展示，用于设置WebIDE资源的规格。请根据实际需求选择具体规格。
实例	当“开发环境”选择“WebIDE”时展示，用于设置当前环境规格对应的环境实例。 ● 如果当前选定的规格有环境实例，可选择已存在的实例。 ● 如果当前选定的规格没有可用的实例，可选择“新建一个新环境”。

步骤3 单击“确定”。

进入模型训练工程详情页面，如图4-86所示。界面介绍如表4-66所示。

图 4-86 模型训练工程详情界面

表 4-66 模型训练工程详情界面说明

区域	参数名称	参数说明
1（训练工程）	创建时间	训练工程创建时间
	类型	模型训练的类型
	创建者	创建训练工程的用户
	活动时间	最近一次模型训练执行的时间
	开发环境	模型训练运行环境信息，可通过下拉框切换当前环境。
		进入模型训练编辑界面
		创建训练任务，详细请参考： • **创建训练任务（简易编辑器）** • **创建训练任务（WebIDE）**
		删除训练工程
	模型训练工程描述	模型训练工程的描述信息，支持单击“  ”编辑描述信息。
		对训练任务的训练报告进行对比，输出训练任务在不同超参下的评估指标，同时显示各训练任务的任务系统参数。 **说明** 最多支持3个模型报告对比。
		切换到其他的训练工程、训练服务或超参优化服务的模型训练页面中。
		Web IDE环境资源配置与管理，包括创建环境、暂停运行中的环境以及删除已有环境。还可查看当前所有配置了Web IDE环境资源的项目的环境信息。

区域	参数名称	参数说明
	创建	新建训练工程、联邦学习工程、训练服务或超参优化服务。
2（模型训练任务）	ALL	根据训练状态快速检索训练任务。
	只展示关注的任务	仅展示关注的任务。 用户可以单击任务名称左侧的  关注指定任务，再次单击  取消关注。
	按任务创建时间	根据任务创建时间、任务名称检索训练任务。 默认按任务创建时间检索。
	倒序	按任务创建时间或者任务名称检索训练任务，检索结果按正序或者倒序排列展示。 默认按倒序排序。
	任务名称	模型训练任务的名称
	任务描述	模型训练任务的描述信息
	任务创建时间	模型训练任务创建的时间
	训练用时	模型训练耗时时长
	Tensorboard	Tensorboard状态
	训练状态	显示训练任务当前的状态。 包括如下状态： • ALL显示所有训练任务。 • WAITING表示训练任务准备中。 • RUNNING表示正在训练。 • FINISHED表示训练成功 • FAILED表示训练失败。 • STOPPED表示停止训练任务。
	评估报告	单击可查看训练评估报告详情。
	资源占用	显示训练算法CPU、GPU和RAM的占用情况。
	峰值	显示训练算法CPU、GPU和RAM使用过程中的峰值。
		训练状态为RUNNING时，可以执行此按钮停止训练任务。
		查看验证任务的详细情况，包括系统日志、运行日志、运行图和Tensorboard。

区域	参数名称	参数说明
		删除训练任务。
		查看优化报告。
		打包训练模型。 **说明** 仅训练成功的模型支持打包。

----结束

4.7.2.2 编辑训练代码（简易编辑器）

编辑代码

支持使用简易编辑器编辑代码，按下“Ctrl+S”保存文件。

可选择下述一种方式，进入简易编辑器开发环境编辑代码：

- 在“模型训练”菜单页面，单击模型训练工程所在行的 。
- 在“模型训练”菜单页面，单击模型训练工程所在行，进入详情界面。单击详情界面右上角的 图标。

简易编辑器界面，如图4-87所示，界面说明如表4-67所示。

图 4-87 简易编辑器界面

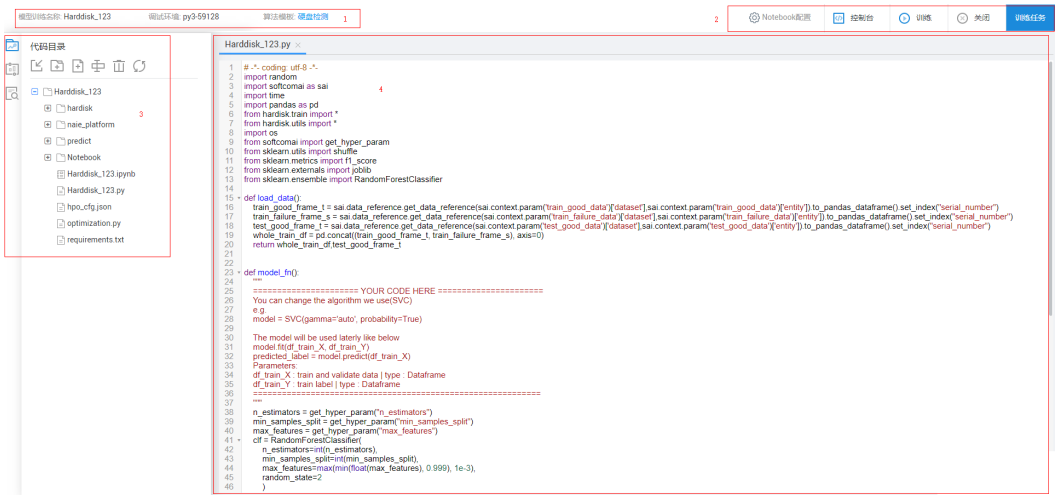
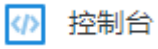
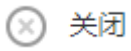
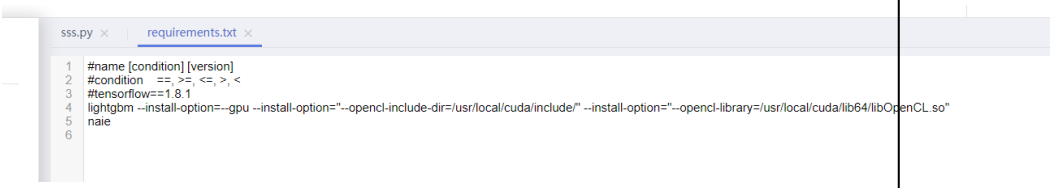


表 4-67 简易编辑器界面说明

区域	说明
1	简易编辑器菜单栏。 ● 模型训练名称：创建模型训练工程时的工程名称。 ● 调试环境：创建调试环境时选择的调试环境。 ● 模型训练模板：使用模板创建项目时显示使用的模板名称。
2	任务执行区。 ●  ：重新配置当前训练工程的调试环境。 ●  ：以页签形式分别显示训练任务的系统日志、运行日志、运行图和Tensorboard。系统支持通过  、  、  刷新、放大及关闭控制台界面；支持通过“Ctrl+F”方式搜索日志。 ●  ：将当前训练工程加入训练。 ●  ：返回到当前训练工程所在的“模型训练”页面。 ● 训练任务：查看训练任务的运行状态。可以查看训练任务的运行日志以及训练报告，删除训练任务。也可以在任务执行过程中单击  暂停训练任务。

区域	说明
3	代码目录：包含日志文件夹、模型文件存放文件夹、调试文件、requirements.txt文件。模型训练/Notebook支持通过requirements.txt安装或升级第三方库。以安装1.0.0版本的pystan为例，操作如下： `pystan == 1.0.0` 且requirements.txt支持带参数的源码安全方式，例如安装GPU版本的lightgbm，如图4-88所示： **图 4-88** 安装 lightgbm  代码目录还支持以下操作： •  ：导入文件。支持上传文件和文件夹两种形式。 •  ：新建文件夹。 •  ：新建文件。 •  ：重命名调试文件、推理文件等文件。 •  ：删除文件或文件夹。 •  ：刷新代码目录。 • 数据集目录：包含数据集文件夹及数据实例。系统支持通过Spread编辑器打开csv文件，支持用户在训练工程编辑界面打开数据集实例。 • 任务目录：包含训练工程已经执行及正在执行的训练任务存储目录结构。包括codes文件、log文件、meta文件、model文件等。
4	代码编辑区。

调试代码

步骤1 单击“Notebook配置”，弹出Notebook配置对话框，配置调试环境。

如果有已经创建好的Notebook环境，直接选中“运行中”的环境，单击“保存”即可。否则需要重新创建Notebook开发环境，操作步骤如下：

1. 从Python版本下拉框中选择指定的Python版本，从调试资源下拉框中选择GPU、CPU调试资源。
2. 单击“创建Notebook环境”。
3. 待环境状态为“运行中”时，选中该环境，单击“保存”。

步骤2 单击“*.ipynb”文件进入调试界面。

步骤3 在弹出的对话框内选择内核，单击“Set Kernel”。

步骤4 在输入框中配置代码，单击 Run 调试代码。

----结束

4.7.2.3 编辑训练代码（WebIDE）

支持使用WebIDE开发环境编辑代码。

可选择下述一种方式，进入WebIDE开发环境编辑代码：

- 在“模型训练”菜单页面，单击模型训练工程所在行的 。其中“开发环境”必须选择WebIDE环境。
- 在“模型训练”菜单页面，单击模型训练工程所在行，进入详情界面。单击详情界面右上角的 图标。其中“开发环境”必须选择WebIDE环境。

WebIDE界面，如图4-89所示，界面说明如表4-68所示。

图 4-89 WebIDE 界面

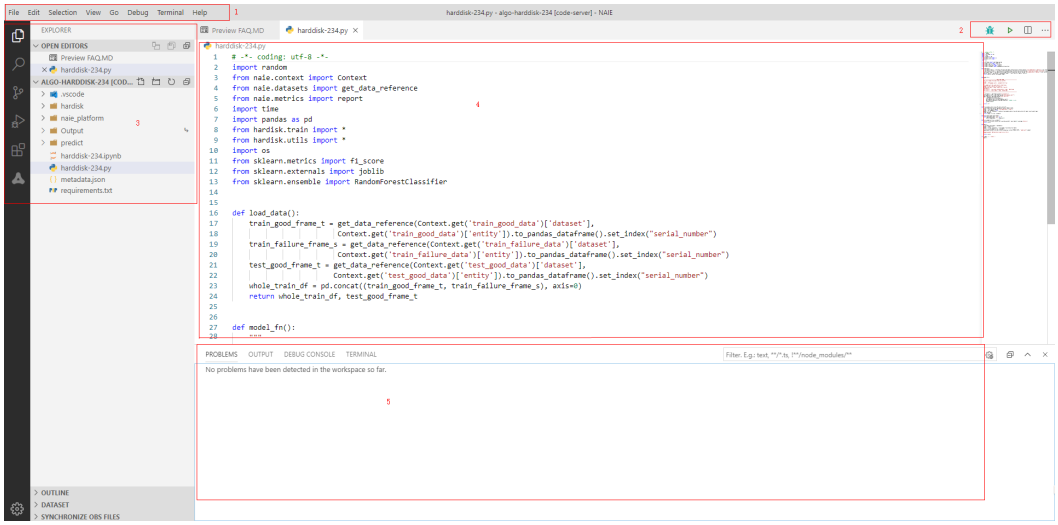


表 4-68 WebIDE 界面说明

区域	说明
1	WebIDE菜单栏。

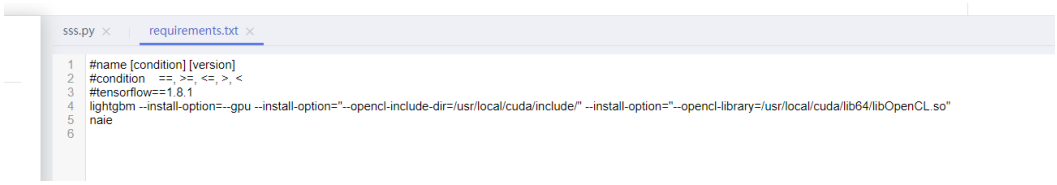
区域	说明
2	代码运行和调试按钮。 •  ：调试代码。 •  ：在终端窗口运行。 •  ：拆分编辑区域，可同时展示多个文件编辑窗口。
3	•  ：文件管理，在文件管理中可以看到所有文件视图，双击文件可在右侧编辑区域编辑。右键单击文件视图空白区域，可打开右键菜单，用户可根据需要使用菜单对应功能。 •  ：查找和替换，输入关键字，在所有文件中查找关键字，并替换关键字。 •  ：git功能，可使用git功能进行版本控制。 •  ：debug面板，调试代码时，可以通过调试面板查看管理变量、堆栈和断点等调试状态。 •  ：插件管理，可以搜索需要的插件并安装，也可以对已安装的插件进行管理，比如卸载、停用等。 •  ：训练任务列表展示，展开训练任务可查看任务下的文件、日志等。
4	代码编辑区。
5	面板区域，从左至右依次为“问题”区域、“输出”区域、“调试”区域和“终端”区域，可以在“终端”区域输入命令行。

其中，代码目录包含日志文件夹、模型文件存放文件夹、调试文件、requirements.txt文件。模型训练/Notebook支持通过requirements.txt安装或升级第三方库。以安装1.0.0版本的pystan为例，操作如下：

```
pystan == 1.0.0
```

且requirements.txt支持带参数的源码安全方式，例如安装GPU版本的lightgbm，如下图所示：

图 4-90 安装 lightgbm



4.7.2.4 模型训练

使用特征工程处理后生成的训练集进行模型训练。

创建训练任务（简易编辑器）

步骤1 单击简易编辑器界面右上角的“训练”，弹出“训练配置”对话框，如图4-91所示。

图 4-91 训练任务配置

步骤2 在“训练配置”对话框中配置参数，如表4-69所示。

表 4-69 训练配置参数配置

区域	参数名称	参数描述
任务说明	任务名称	训练任务的名称。 只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线（_）、（-）组成，不能以下划线结尾，长度范围为[1,32]。
	描述	训练任务的描述信息。
任务运行环境	AI引擎	AI引擎及AI引擎的Python版本。
	创建tensorboard任务	创建Tensorboard，详情请参见创建Tensorboard。
	自定义引擎	通过引擎的镜像地址自定义增加引擎。
	主入口	训练任务的入口文件及入口函数。
	计算节点规格	模型训练服务提供的计算节点资源，包括CPU和GPU。 用户可以单击选定计算节点资源，并在“计算节点个数”中配置计算节点资源的个数。
	计算节点个数	计算节点的个数。 • 1代表单节点计算 • 2代表分布式计算，开发者需要编写相应的调用代码。可使用内置的MoXing分布式训练加速框架进行训练，训练算法需要符合MoXing程序结构。可参考如下文档：https://github.com/huaweicloud/ModelArts-Lab/tree/master/docs/moxing_api_doc

区域	参数名称	参数描述
数据集参数配置	数据集超参	配置数据集实例的超参。 通过调用SDK（get_hyper_param）获取数据集相关的超参，包括训练数据集实例、验证数据集实例等。数据集超参支持输入多个，可以通过“增加”和 图标，来增加或删除运行超参。 详细SDK说明，请在训练服务首页右下角的浮框中，依次单击“帮助中心 > SDK文档”查看。
超参配置	运行超参	通过调用SDK（get_hyper_param）获取运行超参，包括标签列、迭代次数等。运行超参支持输入多个，可以通过“增加”和 图标，来增加或删除运行超参。 详细SDK说明，请在训练服务首页右下角的浮框中，依次单击“帮助中心 > SDK文档”查看。
	超参优化	训练任务执行的过程中可以同步进行超参优化。 勾选“运行超参”后的“超参优化”复选框，可配置运行超参的参数类型、起始值、终止值、优化方法、优化目标和终止条件。训练完成后，可以单击 查看优化报告，得到运行超参不同取值下的模型评分和试验时长。详情请参见**创建超参优化服务**。

步骤3 单击“开始训练”，提交模型训练任务。

注意

如果“训练任务状态”一直处在“RUNNING”中，模型训练服务平台的前台就会一直给后台发消息，查询当前训练任务的状态。即使平台访问超时，查询训练任务状态的接口还是会一直给后台发送查询消息，永不超时。直到“训练任务状态”变更为“FINISHED”、“FAILED”或“STOPPED”，接口才会停止服务状态查询操作。

步骤4 单击 训练任务，查看训练状态。

- ALL显示所有训练任务。
- WAITING表示训练任务准备中。
- RUNNING表示正在训练。
- FINISHED表示训练成功。
- FAILED表示训练失败。
- STOPPED表示停止训练任务。

步骤5 单击训练任务下方的 图标，查看系统日志、运行日志、运行图和Tensorboard信息。

- 系统日志：可以查看代码执行的具体过程。系统运行日志信息，如代码目录、日志路径、使用的SDK信息等。
- 运行日志：用户可以在代码编辑的时候自定义信息输出到运行日志中，用于查看代码执行的具体结果。例如用户信息、代码目录、执行命令等。当训练任务运行失败时，可以通过运行日志定位训练任务失败原因。
- 运行图：用户在训练工程中，调用SDK，以图表的形式显示任务执行信息。
- Tensorboard：创建训练任务时，若勾选了“创建Tensorboard任务”，训练结束后，该页签可以展示TensorFlow在运行过程中的计算图、各种指标随着时间的变化趋势以及训练中使用到的数据信息。

单击图标，查看模型评估报告。

- 评估指标：可以通过数值和图表方式展示各项指标的数据信息。
- 超参：展示训练集、测试集和标签列的信息。
- 任务系统参数：展示训练任务的配置参数信息。

----结束

创建训练任务（WebIDE）

步骤1 返回“模型训练”菜单界面，单击模型训练工程所在行，进入训练工程详情界面。

步骤2 单击界面右上角的图标，弹出“训练任务配置”对话框，如图4-92所示。

图 4-92 训练任务配置

步骤3 在“训练任务配置”对话框中配置参数，如表4-70所示。

表 4-70 训练配置参数配置

区域	参数名称	参数描述
任务说明	任务名称	训练任务的名称。 只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线（_）、（-）组成，不能以下划线结尾，长度范围为[1,32]。
	描述	训练任务的描述信息。
任务运行环境	AI引擎	AI引擎及AI引擎的Python版本。
	创建tensorboard任务	创建Tensorboard，详情请参见创建Tensorboard。
	自定义引擎	通过引擎的镜像地址自定义增加引擎。
	主入口	训练任务的入口文件及入口函数。
	计算节点规格	模型训练服务提供的计算节点资源，包括CPU和GPU。 用户可以单击选定计算节点资源，并在“计算节点个数”中配置计算节点资源的个数。

区域	参数名称	参数描述
	计算节点个数	计算节点的个数。 ● 1代表单节点计算 ● 2代表分布式计算，开发者需要编写相应的调用代码。可使用内置的MoXing分布式训练加速框架进行训练，训练算法需要符合MoXing程序结构。可参考如下文档：https://github.com/huaweicloud/ModelArts-Lab/tree/master/docs/moxing_api_doc
数据集参数配置	数据集超参	配置数据集实例的超参。 通过调用SDK（get_hyper_param）获取数据集相关的超参，包括训练数据集实例、验证数据集实例等。数据集超参支持输入多个，可以通过“增加”和图标，来增加或删除运行超参。 详细SDK说明，请在训练服务首页右下角的浮框中，依次单击“帮助中心 > SDK文档”查看。
超参配置	运行超参	通过调用SDK（get_hyper_param）获取运行超参，包括标签列、迭代次数等。运行超参支持输入多个，可以通过“增加”和图标，来增加或删除运行超参。 详细SDK说明，请在训练服务首页右下角的浮框中，依次单击“帮助中心 > SDK文档”查看。
	超参优化	训练任务执行的过程中可以同步进行超参优化。 勾选“运行超参”后的“超参优化”复选框，可配置运行超参的参数类型、起始值、终止值、优化方法、优化目标和终止条件。训练完成后，可以单击查看优化报告，得到运行超参不同取值下的模型评分和试验时长。详情请参见**创建超参优化服务**。

步骤4 单击“开始训练”，训练任务开始。

步骤5 单击界面右上角的“关闭”，返回模型训练工程详情界面。

“模型训练任务”下方展示新建的训练任务，“训练状态”列展示任务的状态。

- ALL显示所有训练任务。
- WAITING表示训练任务准备中。
- RUNNING表示正在训练。
- FINISHED表示训练成功。
- FAILED表示训练失败。
- STOPPED表示停止训练任务。

> **注意**
>
> 如果“训练任务状态”一直处在“RUNNING”中，模型训练服务平台的前台就会一直给后台发消息，查询当前训练任务的状态。即使平台访问超时，查询训练任务状态的接口还是会一直给后台发送查询消息，永不超时。直到“训练任务状态”变更为“FINISHED”、“FAILED”或“STOPPED”，接口才会停止服务状态查询操作。

步骤6 单击训练任务记录对应的图标，查看系统日志、运行日志、运行图和Tensorboard信息。

- 系统日志：可以查看代码执行的具体过程。系统运行日志信息，如代码目录、日志路径、使用的SDK信息等。
- 运行日志：用户可以在代码编辑的时候自定义信息输出到运行日志中，用于查看代码执行的具体结果。例如用户信息、代码目录、执行命令等。当训练任务运行失败时，可以通过运行日志定位训练任务失败原因。
- 运行图：用户在训练工程中，调用SDK，以图表的形式显示任务执行信息。
- Tensorboard：创建训练任务时，若勾选了“创建Tensorboard任务”，训练结束后，该页签可以展示TensorFlow在运行过程中的计算图、各种指标随着时间的变化趋势以及训练中使用到的数据信息。

单击图标，查看模型评估报告。

- 评估指标：可以通过数值和图表方式展示各项指标的数据信息。
- 超参：展示训练集、测试集和标签列的信息。
- 任务系统参数：展示训练任务的配置参数信息。

----结束

4.7.2.5 MindSpore 样例

MindSpore 是一个全场景 AI 计算框架，它的特性是可以显著减少训练时间和成本（开发态）、以较少的资源和最高能效比运行（运行态），同时适应包括端、边缘与云的全场景（部署态）。

本章介绍如何在训练平台上完成MindSpore样例体验，体验过程中使用的训练算法文件请从NAIE云服务论坛获取。需使用华为云账号登录后，才能下载附件，下载地址如下：

https://bbs.huaweicloud.com/forum/thread-59601-1-1.html

MindSpore体验样例共包含两个算法文件：

- dataset.py

  此算法文件用于加载cifar数据集，并进行简单的数据加强。用户体验MindSpore时，无需进行数据集和特征处理操作。

- resnet.py

  此算法文件为MindSpore体验样例的主入口函数文件，使用MindSpore自带的ResNet50残差网络，并定义了损失函数（SoftmaxCrossEntropyWithLogits）、优化方法（Momentum）、Checkpoint配置，完成网络结构的整体定义。同时作

为主函数，定义了运行超参及其默认值，用户也可以通过训练平台超参配置，覆盖默认值。

步骤1 单击“创建”，弹出“创建训练”对话框。

步骤2 配置MindSpore样例训练工程参数，如图4-93所示。

图 4-93 创建 MindSpore 样例训练工程

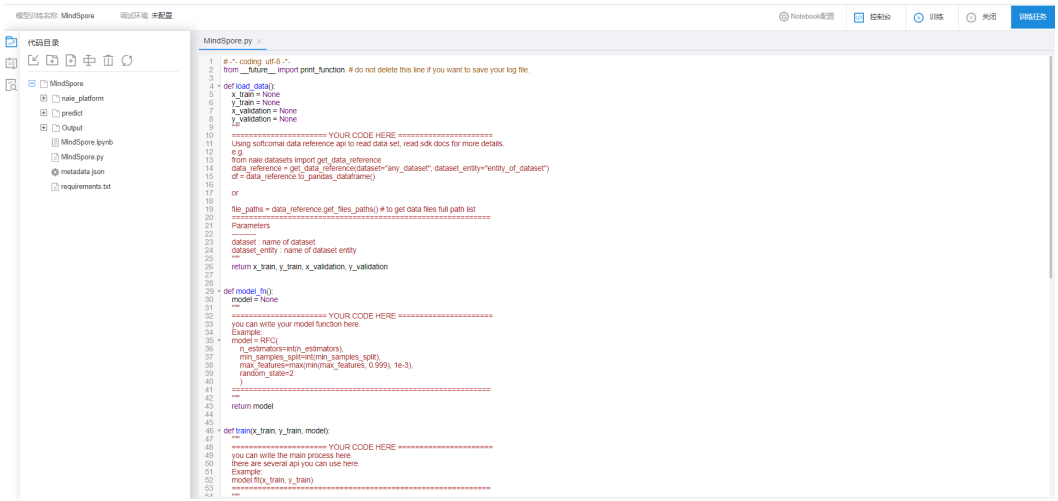
步骤3 单击“确定”。

进入模型训练工程详情界面，如图4-94所示。

图 4-94 模型训练工程详情界面

步骤4 单击界面右上角的图标，进入代码编辑页面，如图4-95所示。

图 4-95 代码编辑页面



步骤5 单击界面左上角的图标，批量上传算法文件，如图4-96所示。

图 4-96 上传算法文件

说明

“resnet.py”文件有两种使用方式：

- 上传至训练工程的代码目录：进行模型训练时，主入口文件选择“resnet.py”。本文采用上传方式描述。
- 不上传至训练工程的代码目录：本地打开算法文件，将该算法文件内容拷贝至与训练工程同名的.py文件中。进行模型训练时，主入口文件选择与训练工程同名的.py文件。

步骤6 单击“上传”。

步骤7 单击界面右上角的“训练”。

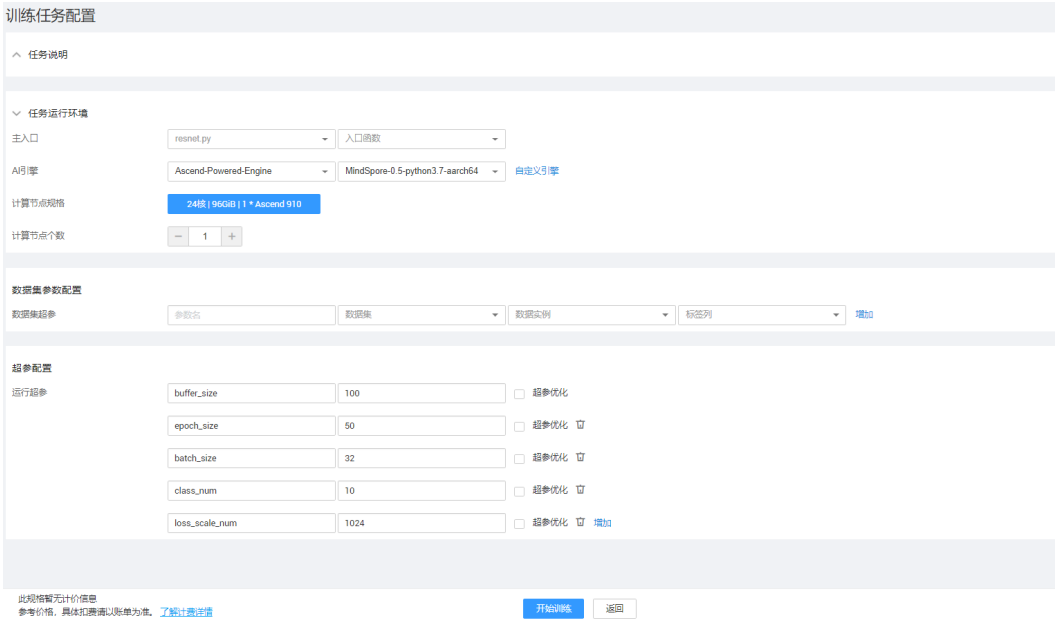
进入“训练任务配置”页面。

步骤8 配置训练任务，如图4-97所示。

参数配置说明如下：

- AI引擎：AI算法运行平台。从第一个下拉框中选择AI引擎“Ascend-Powered-Engine”，从第二个下拉框中选择匹配的python语言版本“MindSpore-0.5-python3.7-aarch64”。
- 主入口：MindSpore样例工程的主算法入口文件，此处选择“resnet.py”。
- 计算节点规格：MindSpore样例模型训练的资源配置信息。
- 计算节点个数：如果配置为“1”，表示使用1个节点进行训练；如果配置为2或者更大，表示使用分布式训练，开发者需要编写相应的调用代码。可使用内置的MoXing分布式训练加速框架进行训练，训练算法需要符合MoXing程序结构。可参考如下文档：https://github.com/huawei-clouds/modelarts-example/blob/master/moxing-apidoc/MoXing-API-UserInstructions.md
- 数据集超参：已预置数据集超参配置，训练任务配置页面无须再配置。
- 运行超参：如图4-97所示为本样例的运行超参，用户可以自行调整超参值，也可以不设置运行超参，使用预置的超参进行训练。

图 4-97 配置 MindSpore 训练任务



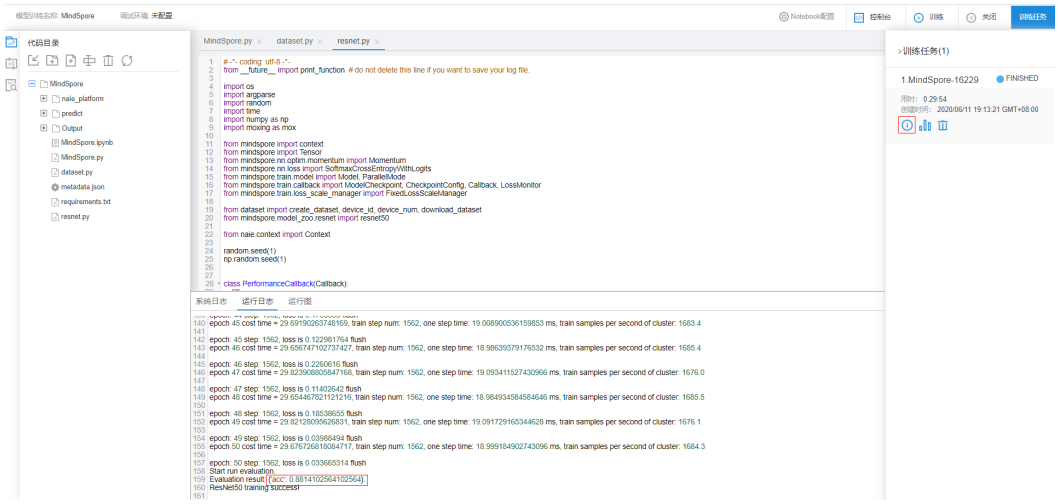
步骤9 单击“开始训练”。

系统返回代码编辑页面。

步骤10 单击界面右上角的“训练任务”，查看训练任务。

待训练状态变为“Finished”时，可单击训练任务下方的图标，查看训练日志，如图4-98所示。“acc”值为模型精度。

图 4-98 查看训练日志



----结束

4.7.3 创建联邦学习工程

4.7.3.1 创建工程

创建联邦学习工程是从创建工程、编辑代码到调试代码的端到端的代码开发过程。

- 创建联邦学习工程：创建联邦学习训练代码编辑和调试的环境。
- 编辑代码：在线编辑联邦学习训练代码。
- 调试代码：在线调试联邦学习训练代码。

创建联邦学习工程步骤如下。

步骤1 单击“创建”，弹出“创建训练”对话框。

配置联邦学习工程参数，如表4-71所示。

表 4-71 参数说明

参数名称	参数说明
请选择模型训练方式	模型训练方式。包含如下选项： ● 新建模型训练工程 ● 新建联邦学习工程 ● 新建训练服务 ● 新建超参优化服务 请选择：新建联邦学习工程。
模型训练名称	模型训练名称。 只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线（_）组成，不能以下划线结尾，长度范围为[1,26]。

参数名称	参数说明
描述	对新建联邦学习工程的描述。
开发环境	训练工程使用的开发环境，支持： • WebIDE WebIDE提供类似本地VSCode的编码体验，支持代码自动补齐、调试等功能，适用于大量代码编写场景。创建在线IDE版联邦学习训练模型时选择“WebIDE”开发环境。 • 简易编辑器 简易编辑器提供代码查看和编辑能力，不支持调试，适用于少量代码修改场景。创建WEB版联邦学习训练模型时，选择“简易编辑器”开发环境。
规格	当“开发环境”选择“WebIDE”时展示，用于设置WebIDE资源的规格。请根据实际需求选择具体规格。
实例	当“开发环境”选择“WebIDE”时展示，用于设置当前环境规格对应的环境实例。 • 如果当前选定的规格有环境实例，可选择已存在的实例。 • 如果当前选定的规格没有可用的实例，可选择“新建一个新环境”。

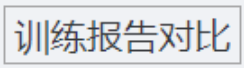
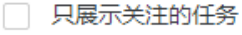
步骤2 单击“确定”。

进入联邦学习工程详情界面，如图4-99所示。界面说明如表4-72所示。

图 4-99 联邦学习工程详情界面

表 4-72 界面说明

区域	参数名称	参数说明
1（训练工程）	创建时间	联邦学习工程创建时间
	类型	模型训练的类型
	创建者	创建联邦学习工程的用户
	活动时间	最近一次模型训练执行的时间
	开发环境	联邦学习模型训练运行环境信息，可通过下拉框切换当前环境。

区域	参数名称	参数说明
		进入代码编辑界面
		创建联邦学习训练任务，详细请参考： • **创建联邦学习训练任务（简易编辑器）** • **创建联邦学习训练任务（WebIDE）**
		删除联邦学习训练工程
	模型训练工程描述	描述信息，支持单击  图标，编辑描述信息。
		对训练任务的训练报告进行对比，输出训练任务在不同超参下的评估指标，同时显示各训练任务的任务系统参数。 **说明** 最多支持3个模型报告对比。
		切换到其他模型训练工程、联邦学习工程、训练服务或超参优化服务详情界面。
		Web IDE环境资源配置与管理，包括创建环境、暂停运行中的环境以及删除已有环境。还可查看当前所有配置了Web IDE环境资源的项目的环境信息。
		新建训练工程、联邦学习工程、训练服务或超参优化服务。
2（模型训练任务）		根据训练状态快速检索训练任务。
		仅展示关注的任务。 用户可以单击任务名称左侧的  关注指定任务，再次单击  取消关注。
		根据任务创建时间、任务名称检索训练任务。 默认按任务创建时间检索。
		按任务创建时间或者任务名称检索训练任务，检索结果按正序或者倒序排列展示。 默认按倒序排序。
	任务名称	模型训练任务的名称
	任务描述	模型训练任务的描述信息
	任务创建时间	模型训练任务创建的时间

区域	参数名称	参数说明
	训练用时	模型训练耗时时长
	Tensorboard	Tensorboard状态
	训练状态	显示训练任务当前的状态。 包括如下状态： ● ALL显示所有训练任务。 ● WAITING表示训练任务准备中。 ● RUNNING表示正在训练。 ● FINISHED表示训练成功 ● FAILED表示训练失败。 ● STOPPED表示停止训练任务。
	评估报告	单击可查看训练评估报告详情。
	资源占用	显示训练算法CPU、GPU和RAM的占用情况。
	峰值	显示训练算法CPU、GPU和RAM使用过程中的峰值。
		训练状态为RUNNING时，可以执行此按钮停止训练任务。
		查看验证任务的详细情况，包括系统日志、运行日志、运行图和Tensorboard。
		删除训练任务。
		查看优化报告。
		打包训练模型。 **说明** 仅训练成功的模型支持打包。

----结束

4.7.3.2 编辑代码（简易编辑器）

编辑代码

支持使用简易编辑器编辑代码。可选择下述一种方式，进入简易编辑器开发环境编辑代码：

- 在“模型训练”菜单页面，单击联邦学习工程所在行的  。

- 在“模型训练”菜单页面，单击联邦学习工程所在行，进入详情界面。单击详情界面右上角的  图标。

简易编辑器界面，如图4-100所示，界面说明如表4-73所示。

图 4-100 简易编辑器界面

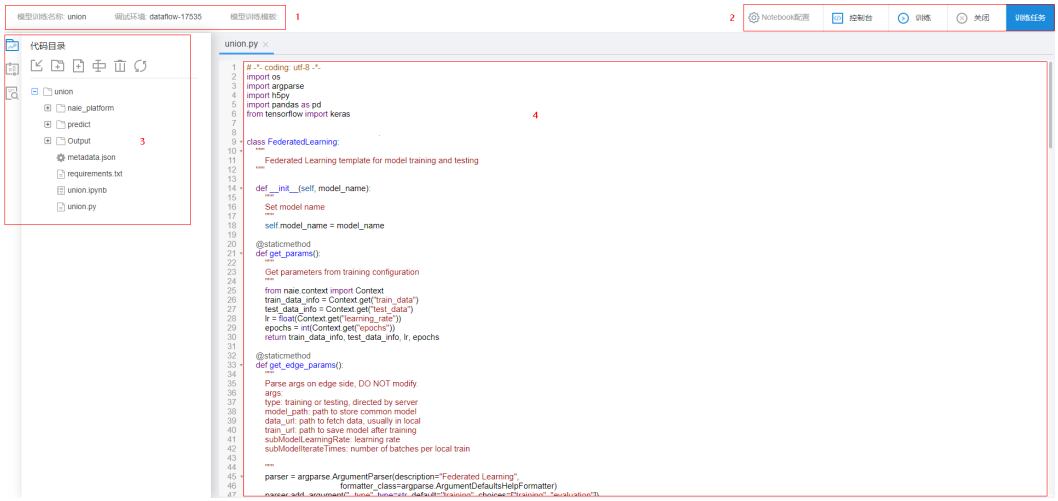


表 4-73 简易编辑器界面说明

区域	说明
1	简易编辑器菜单栏。 ● 模型训练名称：创建模型训练工程时的工程名称。 ● 调试环境：创建调试环境时选择的调试环境。 ● 模型训练模板：使用模板创建项目时显示使用的模板名称。
2	任务执行区。 ● Notebook配置：重新配置当前训练工程的调试环境。 ● 控制台：以页签形式分别显示训练任务的系统日志、运行日志、运行图和Tensorboard。系统支持通过、、刷新、放大及关闭控制台界面；支持通过“Ctrl+F”方式搜索日志。 ● 训练：将当前训练工程加入训练。 ● 关闭：返回到当前训练工程所在的“模型训练”页面。 ● 训练任务：查看训练任务的运行状态。可以查看训练任务的运行日志以及训练报告，删除训练任务。也可以在任务执行过程中单击暂停训练任务。

区域	说明
3	代码目录：包含日志文件夹、模型文件文件夹、调试文件、requirements.txt文件等。模型训练/Notebook支持通过requirements.txt安装或升级第三方库。以安装1.0.0版本的pystan为例，操作如下： `pystan == 1.0.0` 代码目录还支持以下操作： •  ：导入文件。支持上传文件和文件夹两种形式。 •  ：新建文件夹。 •  ：新建文件。 •  ：重命名调试文件、推理文件等文件。 •  ：删除文件或文件夹。 •  ：刷新代码目录。 • 数据集目录：包含数据集文件夹及数据实例。系统支持通过Spread编辑器打开csv文件，支持用户在训练工程编辑界面打开数据集实例。 • 任务目录：包含联邦学习训练工程已经执行及正在执行的训练任务存储目录结构。包括codes文件、log文件、meta文件、model文件等。
4	代码编辑区。

调试代码

步骤1 单击“Notebook配置”，弹出Notebook配置对话框，配置调试环境。

如果有已经创建好的Notebook环境，直接选中“运行中”的环境，单击“保存”即可。否则需要重新创建Notebook开发环境，操作步骤如下：

1. 从Python版本下拉框中选择指定的Python版本，从调试资源下拉框中选择GPU|CPU调试资源。
2. 单击“创建Notebook环境”。
3. 待环境状态为“运行中”时，选中该环境，单击“保存”。

步骤2 单击“*.ipynb”文件进入调试界面。

步骤3 在弹出的对话框内选择内核，单击“Set Kernel”。

步骤4 在输入框中配置代码，单击 

调试代码。

----结束

4.7.3.3 编辑代码（WebIDE）

支持使用WebIDE开发环境编辑代码。可选择下述一种方式，进入WebIDE开发环境编辑代码：

- 在“模型训练”菜单页面，单击联邦学习工程所在行的  。其中“开发环境”必须选择WebIDE环境。
- 在“模型训练”菜单页面，单击联邦学习工程所在行，进入详情界面。单击详情界面右上角的  图标。其中“开发环境”必须选择WebIDE环境。

WebIDE界面，如图4-101所示，界面说明如表4-74所示。

图 4-101 WebIDE 界面

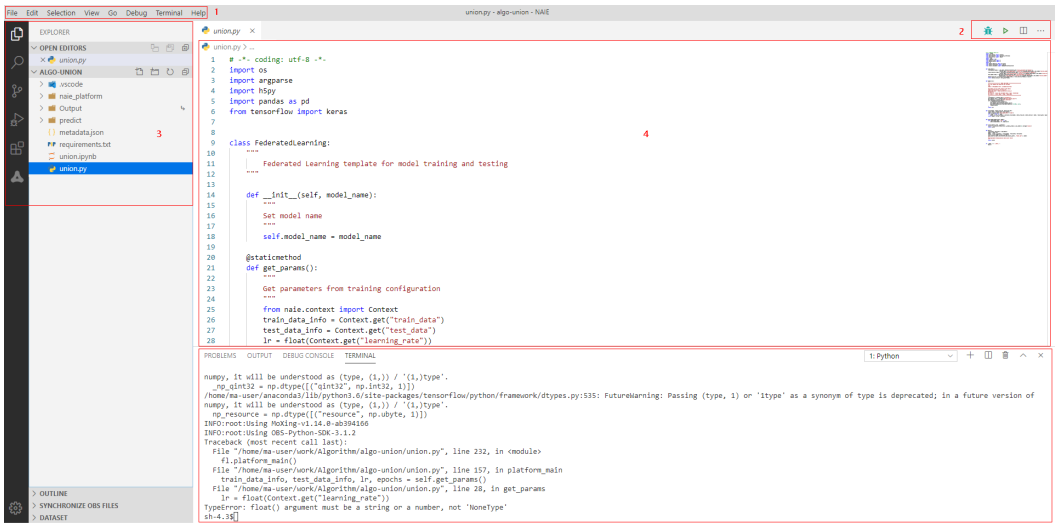


表 4-74 WebIDE 界面说明

区域	说明
1	WebIDE菜单栏。
2	代码运行和调试按钮。 •  ：调试代码。 •  ：在终端窗口运行。 •  ：拆分编辑区域，可同时展示多个文件编辑窗口。

区域	说明
3	•  ：文件管理，在文件管理中可以看到所有文件视图，双击文件可在右侧编辑区域编辑。右键单击文件视图空白区域，可打开右键菜单，用户可根据需要使用菜单对应功能。 •  ：查找和替换，输入关键字，在所有文件中查找关键字，并替换关键字。 •  ：git功能，可使用git功能进行版本控制。 •  ：debug面板，调试代码时，可以通过调试面板查看管理变量、堆栈和断点等调试状态。 •  ：插件管理，可以搜索需要的插件并安装，也可以对已安装的插件进行管理，比如卸载、停用等。 •  ：训练任务列表展示，展开训练任务可查看任务下的文件、日志等。
4	代码编辑区。
5	面板区域，从左至右依次为“问题”区域、“输出”区域、“调试”区域和“终端”区域，可以在“终端”区域输入命令行。

4.7.3.4 模型训练

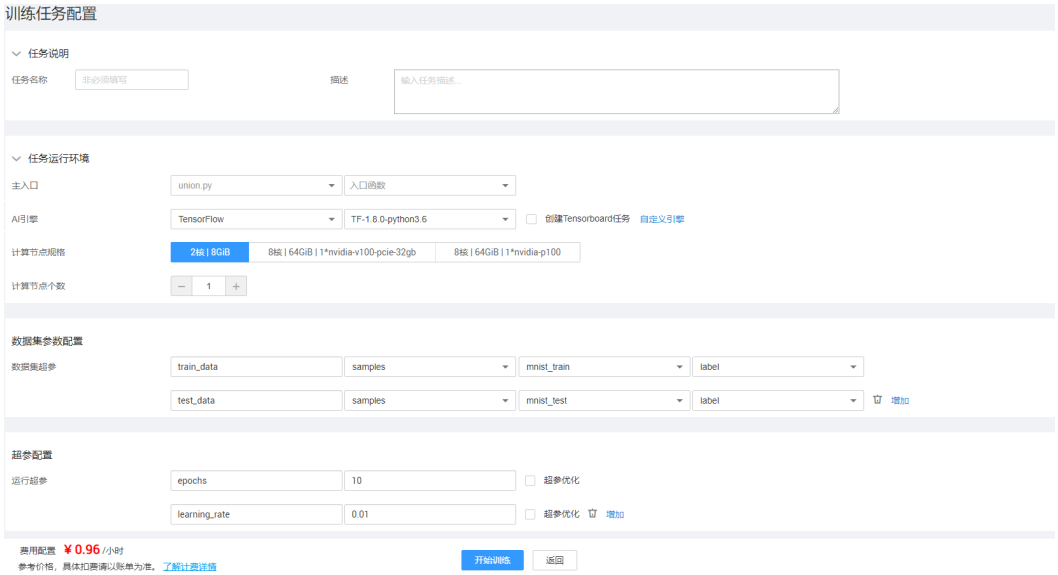
使用特征工程处理后生成的训练集进行模型训练。

创建联邦学习训练任务（简易编辑器）

步骤1 单击简易编辑器界面右上角的“训练”。

进入“训练任务配置”界面，如图4-102所示。

图 4-102 训练任务配置



参数说明，如**表4-75**所示。

表 4-75 参数配置

区域	参数名称	参数描述
任务说明	任务名称	训练任务的名称。 只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线（_）、（-）组成，不能以下划线结尾，长度范围为[1,32]。
	描述	训练任务的描述信息。
任务运行环境	AI引擎	AI引擎及AI引擎的Python版本。
	创建tensorboard任务	创建Tensorboard，详情请参见**创建Tensorboard**。
	自定义引擎	通过引擎的镜像地址自定义增加引擎。
	主入口	训练任务的入口文件及入口函数。
	计算节点规格	模型训练服务提供的计算节点资源，包括CPU和GPU。 用户可以单击选定计算节点资源，并在“计算节点个数”中配置计算节点资源的个数。

区域	参数名称	参数描述
	计算节点个数	计算节点的个数。 ● 1代表单节点计算 ● 2代表分布式计算，开发者需要编写相应的调用代码。可使用内置的MoXing分布式训练加速框架进行训练，训练算法需要符合MoXing程序结构。可参考如下文档：https://github.com/huaweicloud/ModelArts-Lab/tree/master/docs/moxing_api_doc
数据集参数配置	数据集超参	配置数据集实例的超参。 通过调用SDK（get_hyper_param）获取数据集相关的超参，包括训练数据集实例、测试数据集实例等。数据集超参支持输入多个，可以通过“增加”和图标，来增加或删除运行超参。 详细SDK说明，请在训练服务首页右下角的浮框中，依次单击“帮助中心 > SDK文档”查看。
超参配置	运行超参	通过调用SDK（get_hyper_param）获取运行超参，包括标签列、迭代次数等。运行超参支持输入多个，可以通过“增加”和图标，来增加或删除运行超参。 详细SDK说明，请在训练服务首页右下角的浮框中，依次单击“帮助中心 > SDK文档”查看。
	超参优化	训练任务执行的过程中可以同步进行超参优化。 勾选“运行超参”后的“超参优化”复选框，可配置运行超参的参数类型、起始值、终止值、优化方法、优化目标和终止条件。训练完成后，可以单击查看优化报告，得到运行超参不同取值下的模型评分和试验时长。详情请参见**创建超参优化服务**。

步骤2 单击“开始训练”，训练任务开始。

注意

如果“训练任务状态”一直处在“RUNNING”中，模型训练服务平台的前台就会一直给后台发消息，查询当前训练任务的状态。即使平台访问超时，查询训练任务状态的接口还是会一直给后台发送查询消息，永不超时。直到“训练任务状态”变更为“FINISHED”、“FAILED”或“STOPPED”，接口才会停止服务状态查询操作。

步骤3 单击 训练任务，查看训练状态。

- ALL显示所有训练任务。
- WAITING表示训练任务准备中。
- RUNNING表示正在训练。

- FINISHED表示训练成功。
- FAILED表示训练失败。
- STOPPED表示停止训练任务。

步骤4 单击训练任务下方的图标，查看系统日志、运行日志、运行图和Tensorboard信息。

- 系统日志：可以查看代码执行的具体过程。系统运行日志信息，如代码目录、日志路径、使用的SDK信息等。
- 运行日志：用户可以在代码编辑的时候自定义信息输出到运行日志中，用于查看代码执行的具体结果。例如用户信息、代码目录、执行命令等。当训练任务运行失败时，可以通过运行日志定位训练任务失败原因。
- 运行图：用户在训练工程中，调用SDK，以图表的形式显示任务执行信息。
- Tensorboard：创建训练任务时，若勾选了“创建Tensorboard任务”，训练结束后，该页签可以展示TensorFlow在运行过程中的计算图、各种指标随着时间的变化趋势以及训练中使用到的数据信息。

单击图标，查看模型评估报告。

- 评估指标：可以通过数值和图表方式展示各项指标的数据信息。
- 超参：展示训练集、测试集和标签列的信息。
- 任务系统参数：展示训练任务的配置参数信息。

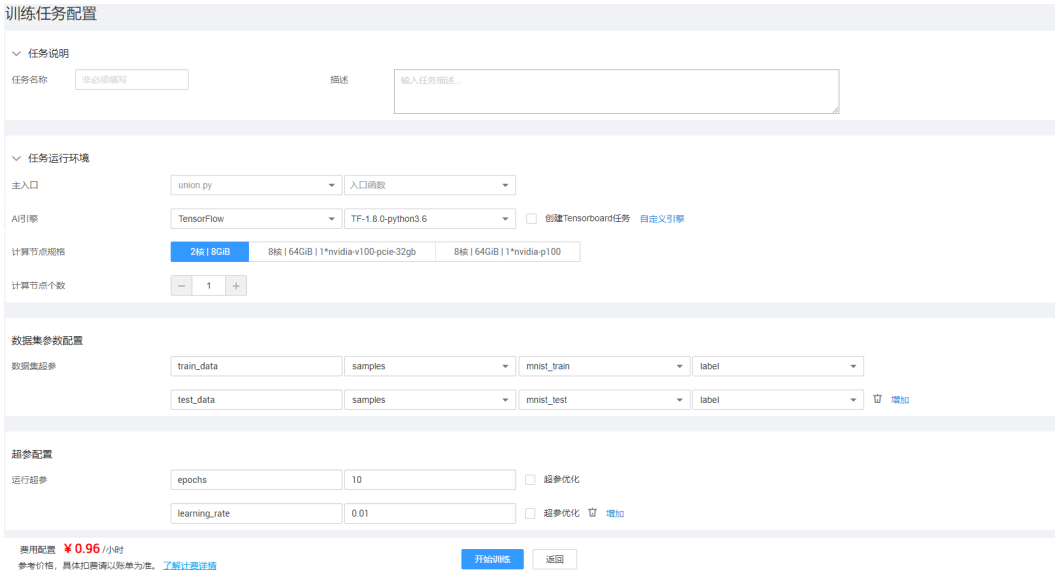
----结束

创建联邦学习训练任务（WebIDE）

步骤1 返回“模型训练”菜单界面，单击联邦学习工程所在行，进入工程详情界面。

步骤2 单击界面右上角的，弹出“训练任务配置”对话框，如图4-103所示。

图 4-103 训练任务配置



参数说明，如表4-76所示。

表 4-76 参数说明

区域	参数名称	参数描述
任务说明	任务名称	训练任务的名称。 只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线（_）、（-）组成，不能以下划线结尾，长度范围为[1,32]。
	描述	训练任务的描述信息。
任务运行环境	AI引擎	AI引擎及AI引擎的Python版本。
	创建tensorboard任务	创建Tensorboard，详情请参见创建Tensorboard。
	自定义引擎	通过引擎的镜像地址自定义增加引擎。
	主入口	训练任务的入口文件及入口函数。
	计算节点规格	模型训练服务提供的计算节点资源，包括CPU和GPU。 用户可以单击选定计算节点资源，并在“计算节点个数”中配置计算节点资源的个数。
	计算节点个数	计算节点的个数。 ● 1代表单节点计算 ● 2代表分布式计算，开发者需要编写相应的调用代码。可使用内置的MoXing分布式训练加速框架进行训练，训练算法需要符合MoXing程序结构。可参考如下文档：https://github.com/huaweicloud/ModelArts-Lab/tree/master/docs/moxing_api_doc
数据集参数配置	数据集超参	配置数据集实例的超参。 通过调用SDK（get_hyper_param）获取数据集相关的超参，包括训练数据集实例、验证数据集实例等。数据集超参支持输入多个，可以通过“增加”和图标，来增加或删除运行超参。 详细SDK说明，请在训练服务首页右下角的浮框中，依次单击“帮助中心 > SDK文档”查看。
超参配置	运行超参	通过调用SDK（get_hyper_param）获取运行超参，包括标签列、迭代次数等。运行超参支持输入多个，可以通过“增加”和图标，来增加或删除运行超参。 详细SDK说明，请在训练服务首页右下角的浮框中，依次单击“帮助中心 > SDK文档”查看。

区域	参数名称	参数描述
	超参优化	训练任务执行的过程中可以同步进行超参优化。 勾选“运行超参”后的“超参优化”复选框，可配置运行超参的参数类型、起始值、终止值、优化方法、优化目标和终止条件。训练完成后，可以单击  查看优化报告，得到运行超参不同取值下的模型评分和试验时长。详情请参见**创建超参优化服务**。

步骤3 单击“开始训练”，训练任务开始。

步骤4 单击“关闭”，返回联邦学习工程详情界面，“模型训练任务”下方展示新建的训练任务，“训练状态”列展示任务的状态。

- ALL显示所有训练任务。
- WAITING表示训练任务准备中。
- RUNNING表示正在训练。
- FINISHED表示训练成功。
- FAILED表示训练失败。
- STOPPED表示停止训练任务。

> **⚠ 注意**
>
> 如果“训练任务状态”一直处在“RUNNING”中，模型训练服务平台的前台就会一直给后台发消息，查询当前训练任务的状态。即使平台访问超时，查询训练任务状态的接口还是会一直给后台发送查询消息，永不超时。直到“训练任务状态”变更为“FINISHED”、“FAILED”或“STOPPED”，接口才会停止服务状态查询操作。

步骤5 单击训练任务所在行的  图标，查看系统日志、运行日志、运行图和Tensorboard信息。

- 系统日志：可以查看代码执行的具体过程。系统运行日志信息，如代码目录、日志路径、使用的SDK信息等。
- 运行日志：用户可以在代码编辑的时候自定义信息输出到运行日志中，用于查看代码执行的具体结果。例如用户信息、代码目录、执行命令等。当训练任务运行失败时，可以通过运行日志定位训练任务失败原因。
- 运行图：用户在训练工程中，调用SDK，以图表的形式显示任务执行信息。
- Tensorboard：创建训练任务时，若勾选了“创建Tensorboard任务”，训练结束后，该页签可以展示TensorFlow在运行过程中的计算图、各种指标随着时间的变化趋势以及训练中使用到的数据信息。

单击  图标，查看模型评估报告。

- 评估指标：可以通过数值和图表方式展示各项指标的数据信息。
- 超参：展示训练集、测试集和标签列的信息。

- 任务系统参数：展示训练任务的配置参数信息。

----结束

4.7.4 创建训练服务

新建训练服务

训练任务需要基于已经成功打包的训练模型去创建，并选择新的训练数据集、测试数据集和标签列进行模型训练。

步骤1 单击“创建”，弹出“创建训练”对话框。

配置训练服务参数，如**新建算法参数说明**所示。

表 4-77 参数说明

参数名称	参数说明
请选择模型训练方式	模型训练方式，包含如下选项： ● 新建模型训练工程 ● 新建联邦学习工程 ● 新建训练服务 ● 新建超参优化服务 请选择：新建训练服务。
描述	对新建训练服务的描述信息。
训练服务名称	训练服务名称。 只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线（_）组成，不能以下划线结尾，长度范围为[1,26]。
归档模型包	从下拉框中选择已归档的模型。

步骤2 单击“确定”。

进入训练服务详情界面，如**图4-104**所示。界面说明如**表4-78**所示。

图 4-104 模型训练

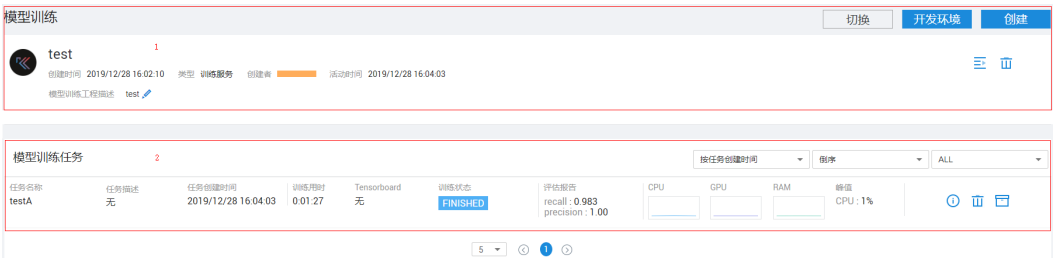


表 4-78 界面说明

区域	参数名称	参数说明
1（训练服务）	创建时间	训练服务创建时间。
	类型	模型训练的类型。
	创建者	创建训练服务的用户。
	活动时间	最近一次模型训练执行的时间。
		创建训练任务，详细请参考**模型训练**。
		删除训练任务。
	模型训练工程描述	训练服务的描述信息，支持单击“  ”重新编辑。
		切换到其他的训练工程、联邦学习工程、训练服务或超参优化服务的模型训练页面中。
		模型训练运行环境信息查看和配置。
		新建训练工程、联邦学习工程、训练服务或超参优化服务。
2（模型训练任务）		根据训练状态快速检索训练任务。
		根据任务创建时间、任务名称检索训练任务。 默认按任务创建时间检索。
		按任务创建时间或者任务名称检索训练任务，检索结果按正序或者倒序排列展示。 默认按倒序排序。
	任务名称	模型训练任务的名称。
	任务描述	模型训练任务的描述信息
	任务创建时间	模型训练任务创建的时间。
	训练用时	模型训练耗时时长。
	Tensorboard	Tensorboard状态。

区域	参数名称	参数说明
	训练状态	显示训练任务当前的状态。 包括如下状态： • ALL显示所有训练任务。 • WAITING表示训练任务准备中。 • RUNNING表示正在训练。 • FINISHED表示训练成功 • FAILED表示训练失败。 • STOPPED表示被停止的训练任务。
	评估报告	单击可查看训练评估报告详情。
	资源占用	显示训练算法CPU、GPU和RAM的占用情况。
	峰值	显示训练算法CPU、GPU和RAM使用过程中的峰值。
		查看训练任务的系统日志、运行日志和运行图。
		训练状态为RUNNING时，可以执行此按钮停止训练任务。
		删除训练任务。
		打包训练模型。 **说明** 仅训练成功的模型支持打包。

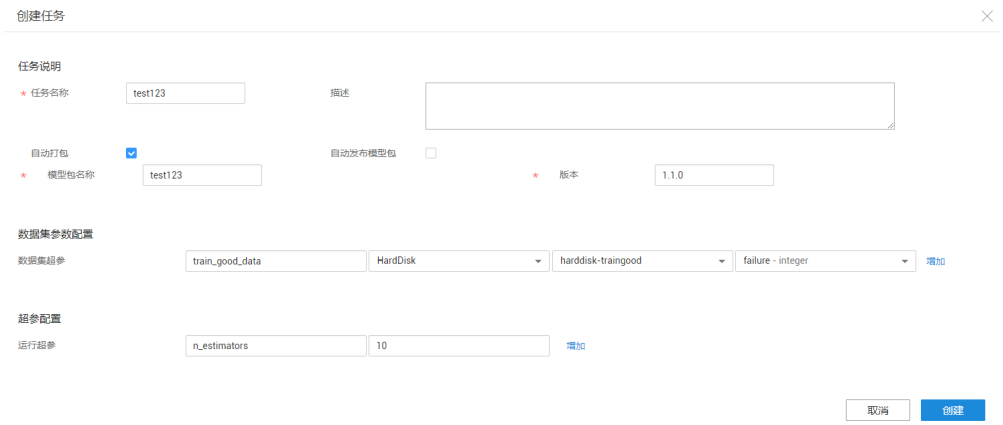
----结束

模型训练

步骤1 单击训练服务详情界面的  图标。

弹出“创建任务”对话框，如图4-105所示。

图 4-105 创建任务



参数说明如**表4-79**所示。

表 4-79 参数说明

区域	参数名称	参数描述
任务说明	任务名称	模型训练任务的名称。 只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线（_）（-）组成，不能以下划线结尾，长度范围为[1,32]。
	描述	对任务的描述信息。
	自动打包	勾选后，创建模型训练任务的同时打包该模型。任务创建成功后可在“模型管理”界面看到打包的模型。
	自动发布模型包	勾选“自动打包”才会展示该参数。勾选“自动发布模型包”，创建模型训练任务的同时打包该模型，并且将打包的模型自动上架。任务创建成功后可在“模型管理”界面看到“上架状态”为“上架中”的模型。
	模型包名称	勾选“自动打包”才会展示该参数，表示模型包打包名称。
	版本	勾选“自动打包”才会展示该参数，表示模型包打包版本。
数据集参数配置	数据集超参	设置当前训练任务的数据集超参，与**模型训练**保持一致。
超参配置	运行超参	运行超参的名称，与**模型训练**保持一致。

步骤2 单击“创建”，训练任务开始。

步骤3 单击 查看任务运行的详细情况，包括系统日志、运行日志和运行图。在评估报告中查看训练结果。

----结束

4.7.5 创建超参优化服务

超参优化服务可以对已创建好的模型训练工程进行超参调优，通过训练结果对比，选择一组最优超参组合。并不是所有的训练工程都可以创建超参优化服务。创建超参优化服务对已创建的训练工程要求如下：

- 训练工程是可以成功执行训练任务的
- 训练工程中超参是通过SDK（softcomai.get_hyper_param）调用的，不在训练代码中定义取值。
- 训练工程需要反馈优化程序所需要的分数

详细的超参优化服务，请参见SDK文档最新版本的“超参数优化示例”。SDK说明请在训练服务首页右下角的浮框中，依次单击“帮助中心 > SDK文档”查看。

新建超参优化服务

超参优化服务为已经创建的模型训练工程，通过训练结果对比，选择一组最优超参组合。

步骤1 单击“创建”，弹出“创建训练”对话框。

参数说明，如表4-80所示。

表 4-80 参数说明

参数名称	参数说明
请选择模型训练方式	模型训练方式。 包含如下选项： • 新建模型训练工程 • 新建联邦学习工程 • 新建训练服务 • 新建超参优化服务 请选择：新建超参优化服务。
描述	描述信息。
优化服务名称	训练服务名称。 只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线（_）组成，不能以下划线结尾，长度范围为[1,26]。
目标训练工程	已经创建的模型训练工程。训练工程创建请参见创建模型训练工程。

步骤2 单击“确定”。

进入超参优化服务详情界面，如图4-106所示。界面说明如表4-81所示。

图 4-106 超参优化服务详情界面

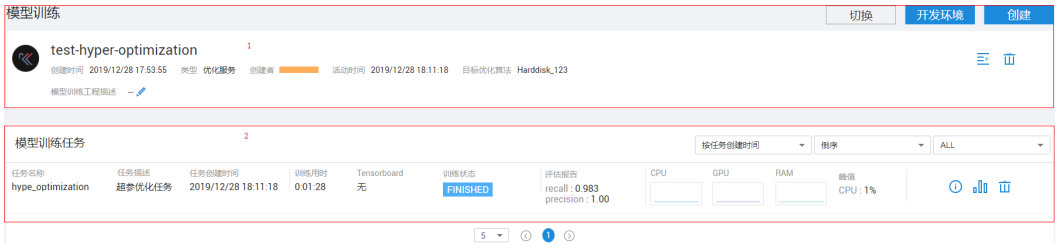


表 4-81 界面说明

区域	参数名称	参数说明
1（训练服务）	创建时间	超参优化服务创建时间。
	类型	模型训练的类型。
	创建者	创建超参优化服务的用户。
	活动时间	最近一次模型训练执行的时间。
	目标优化算法	创建超参优化服务时选择的目标训练工程。
	模型训练工程描述	超参优化服务的描述信息，支持通过单击“  ”重新编辑。
		创建训练任务，详细请参考**模型训练**。
		删除训练任务。
		切换到其他的训练工程、联邦学习工程、训练服务或超参优化服务的模型训练页面中。
		模型训练运行环境信息查看和配置。
		新建训练工程、联邦学习工程、训练服务或超参优化服务。
2（模型训练任务）		根据训练状态快速检索训练任务。
		根据任务创建时间、任务名称检索训练任务。 默认按任务创建时间检索。

区域	参数名称	参数说明
	倒序	按任务创建时间或者任务名称检索训练任务，检索结果按正序或者倒序排列展示。 默认按倒序排序。
	任务名称	模型训练任务的名称。
	任务描述	模型训练任务的描述信息
	任务创建时间	模型训练任务创建的时间。
	训练用时	模型训练耗时时长。
	Tensorboard	Tensorboard状态。
	训练状态	显示训练任务当前的状态。 包括如下状态： • ALL显示所有训练任务。 • WAITING表示训练任务准备中。 • RUNNING表示正在训练。 • FINISHED表示训练成功 • FAILED表示训练失败。 • STOPPED表示被停止的训练任务。
	评估报告	单击可查看训练评估报告详情。
	资源占用	显示训练算法CPU、GPU和RAM的占用情况。
	峰值	显示训练算法CPU、GPU和RAM使用过程中的峰值。
		查看训练任务的系统日志、运行日志和运行图。
		查看优化报告。
		训练状态为RUNNING时，可以执行此按钮停止训练任务。
		删除训练任务。

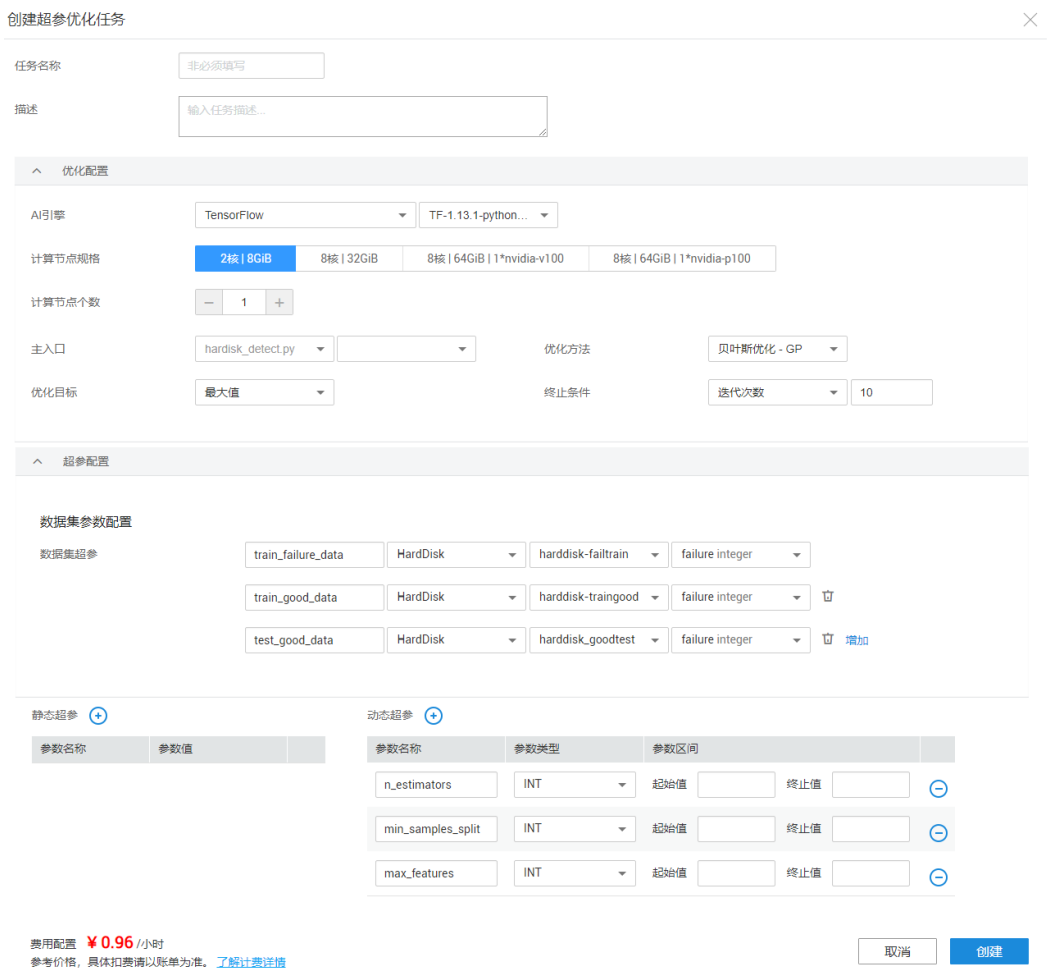
----结束

模型训练

步骤1 在超参优化服务详情界面，单击界面右上角的  图标。

弹出“创建超参优化任务”对话框，如图4-107所示。

图 4-107 创建超参优化任务



参数说明，如创建超参优化任务参数说明所示。

表 4-82 参数说明

区域	参数名称	参数描述
任务名称	任务名称	模型训练任务的名称。
描述	描述	模型训练任务的描述信息。
优化配置	AI引擎	AI引擎及AI引擎的Python版本。

区域	参数名称	参数描述
	计算节点规格	计算节点规格。 模型训练服务提供的计算节点资源，包括CPU和GPU。 用户可以单击选定计算节点资源，并在“计算节点个数”中配置计算节点资源的个数。
	计算节点个数	计算节点的个数： ● 1代表单节点计算 ● 2代表分布式计算
	主入口	训练任务的入口文件及入口函数。
	优化方法	超参优化方法： ● 贝叶斯优化-GP ● 贝叶斯优化-SMAC ● 贝叶斯优化-TPE ● 随机搜索 ● 网格搜索
	优化目标	超参优化任务的目标，在训练算法中进行定义并反馈。根据训练代码选择“最大值”或“最小值”。
	终止条件	● 迭代次数 ● 时间
超参配置	数据集超参	配置数据集实例的超参。 通过调用SDK（get_hyper_param）获取数据集相关的超参，包括训练数据集实例、验证数据集实例等。数据集超参支持输入多个，可以通过“增加”或图标，来增加或删除数据集超参。 详细SDK说明，请在训练服务首页右下角的浮框中，依次单击“在线帮助 > SDK文档”查看。
	静态超参	每次迭代训练，超参的取值是固定不变的： ● 参数名称：静态超参的名称 ● 参数值：静态超参的取值 通过调用SDK（get_hyper_param）获取静态超参，可以通过、图标，来增加或删除静态超参。 详细SDK说明，请在训练服务首页右下角的浮框中，依次单击“帮助中心 > SDK文档”查看。

区域	参数名称	参数描述
	动态超参	每次迭代训练，超参的取值都会依据优化方法重新赋值： ● 参数名称：动态超参的名称 ● 参数类型：动态超参的类型，例如INT、FLOAT、STRING、BOOL等 ● 参数区间：动态超参的取值范围，[起始值，终止值] 通过调用SDK（get_hyper_param）获取动态超参，可通过  增加或删除动态超参。 详细SDK说明，请在训练服务首页右下角的浮框中，依次单击“帮助中心 > SDK文档”查看。

步骤2 单击“创建”，训练任务开始。

步骤3 单击  查看训练任务执行的详细情况，包括系统日志、运行日志和运行图。

----结束

超参优化任务结果查看

在“模型训练”页面，单击  可以查看超参优化任务的优化报告。优化报告中包含：

- 超参优化任务的详细信息：最优超参组合的模型评分、训练耗时、参数取值，以及超参优化任务的参数信息。
- 评分图：在图表中显示每次迭代训练得到的模型评分。
- 超参图：在图表中显示每次迭代训练的超参取值及对应的模型评分。
- 试验时长图：在图表中显示每次迭代训练的超参取值及对应的训练时长。

4.7.6 创建 Tensorboard

TensorBoard是一个可视化工具，能够有效地展示TensorFlow在运行过程中的计算图、各种指标随着时间的变化趋势以及训练中使用到的数据信息。TensorBoard当前只支持基于TensorFlow引擎的训练作业。同一个用户的多个项目，创建Tensorboard任务数不能超过5个。TensorBoard相关概念请参考**TensorBoard官网**。

对于采用AI引擎为TensorFlow的训练作业，您可以使用模型训练时产生的Summary文件来创建TensorBoard作业，将需要展示的指标及数据信息写入到Context.get("tensorboard_path")路径下，示例代码如下：

```
import tensorflow as tf
from naie.context import Context
with tf.name_scope('graph') as scope:
    matrix1 = tf.constant([[3., 3.]],name ='matrix2')
    matrix2 = tf.constant([[2.],[2.]],name ='matrix3')
    product = tf.matmul(matrix1, matrix2,name='product')
sess = tf.Session()
writer = tf.summary.FileWriter(Context.get("tensorboard_path"), sess.graph)
init = tf.global_variables_initializer()
sess.run(init)
```


创建Tensorboard有三个入口：

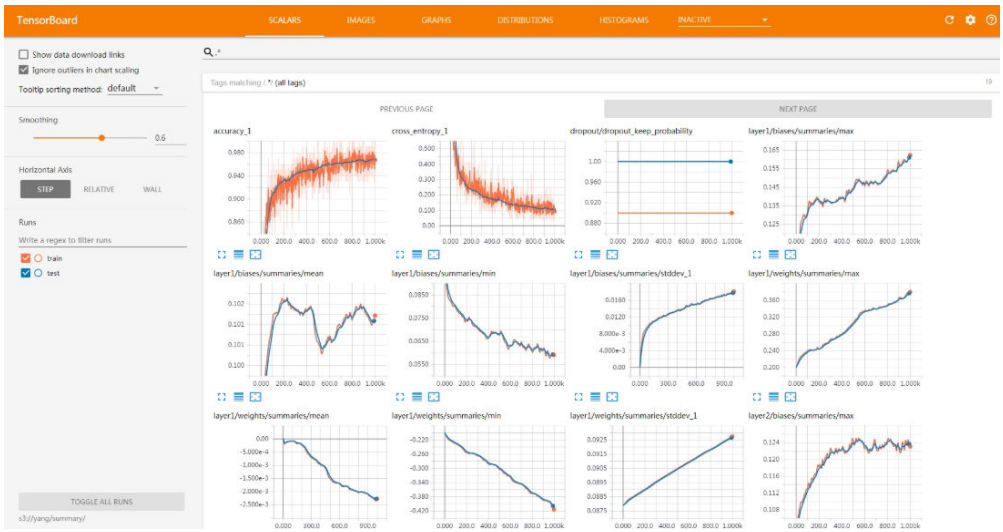
- 创建训练任务的时候同步创建Tensorboard
- 在模型训练工程代码编辑界面控制台的Tensorboard页签中创建Tensorboard
- 新建模型训练工程，创建训练任务后，在任务详情的Tensorboard页签中创建Tensorboard

此处以在训练任务详情的Tensorboard页签中创建Tensorboard为例进行介绍，操作步骤如下。

步骤1 在新建模型训练工程的训练任务界面中，单击训练任务右侧的，进入训练任务详情页面。

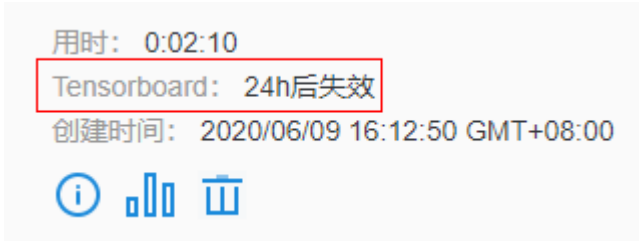
步骤2 在训练任务详情页面，选择“Tensorboard”页签，单击“创建”，完成TensorBoard任务的创建。如图4-108所示。

图 4-108 TensorBoard 界面



TensorBoard任务创建后，训练任务界面新增TensorBoard状态展示，如图4-109所示。

图 4-109 TensorBoard 状态



步骤3 单击页面上方账号信息区域，在下拉菜单中选择“TensorBoard”，可对当前创建的TensorBoard环境进行管理，如删除TensorBoard环境，以及单击环境名，跳转到相应训练任务。

----结束

4.7.7 打包训练模型

系统支持将训练好的模型归档以及打包成模型包。用户可以基于模型包创建验证服务、训练服务。模型验证服务详情可以在**模型验证**查看。模型训练服务详情可以在**创建训练服务**查看。

模型包主要包括模型验证服务的推理主入口函数、特征工程操作流、模型文件等。已发布的模型可以在**模型管理**查看。

本章节介绍的模型打包适用于单个模型打包场景。如果用户需要将多个模型打成一个模型包，或者当用户需要引入外部模型文件时，可以在“模型管理”页面中使用“新建模型包”功能，具体操作请参见**创建模型包**。

步骤1 单击模型训练任务对应的  ，弹出“归档”对话框。

说明

仅支持对训练成功的模型打包，可重复打包。

步骤2 在“归档”对话框中配置参数，如**表4-83**所示。

表 4-83 参数配置

参数名称	参数说明
归档名	归档模型的包名。
归档版本	归档训练模型的版本。 默认版本为1.0.0。
生成模型包	是否直接在归档的同时打包模型包。 选择“是”，表示同时对模型执行归档和打包操作；选择“否”表示仅对模型执行归档操作。默认选择“是”。
包含代码	模型包是否包含训练和推理的相关代码。 选择“是”，表示包含，选择“否”，表示不包含。默认选择“是”。
模型描述	训练模型的描述信息。

步骤3 单击“确定”。

用户可以在**模型管理**对训练模型进行管理操作。

----结束

4.8 模型管理

4.8.1 模型管理简介

训练模型的开发和调优往往需要大量的迭代和调试，数据集的变化、训练算法或者超参的变化都可能会影响模型的质量。用户可将训练完成的优质模型打包到模型管理中，进行统一管理。模型管理中可以查看模型包的详细信息、将多个归档好或者打包

好的模型合打成一个模型包、发布模型包至应用市场、创建联邦学习实例、发布成在线推理服务。

具体操作请参见表4-84。

表 4-84 模型管理操作

参数名称	参数描述
模型名称	模型的名称，与模型打包时保持一致。
模型版本	模型的版本，与模型打包时保持一致。
模型描述	模型的描述内容，与模型打包时保持一致。 不能超过256个字符。
上架状态	模型包的发布状态： ● 未上架：未提交上架。 ● 上架中：提交上架成功，等待应用市场审批。 ● 上架成功：已发布到应用市场。 ● 上架失败：发布到应用市场失败。
创建时间	模型训练任务的打包的时间。
更新时间	模型包最近一次更新时间。
开发环境	模型包运行的开发环境。

参数名称	参数描述
操作	•  ：编辑模型包。编辑模型包内的代码文件，上传新文件等。当前模型包配置了开发环境时，才可以对模型包进行编辑。 •  ：下载模型包。 •  ：将模型包上架到应用市场。 •  ：发布成推理服务。配置请参见**发布推理服务**。 •  ：发布推理服务成功后，可通过此图标进入推理服务的快速验证界面。 •  ：推理服务发布失败，单击可重新发布推理服务。 •  ：已发布推理服务的模型包更新后，单击可更新发布推理服务，更新推理服务版本号最后一位默认在原版本基础上加1。 •  ：创建联邦学习实例。创建步骤请参见**https://support.huawei.com/carrierics/Model%20Training%20%26%20Domain%20Model/Latest%20Version/topic/view.do?portalid=1575625982546&hdxfileid=DOC29856&pidid=pid_bookmap_0189622602&topicid=TOPIC_0208331303&relationid=default&path=DOCNAVI0ED2C09A97B4472EBF80C40BD0DB945B** •  ：删除模型包。 •  ：生成完整性校验SHA256校验码。

4.8.2 创建模型包

“新建模型包”适用于多模型打包场景，用户可以将多个归档的模型合打成一个模型包。或者当用户需要引入外部模型文件时，可以使用“新建模型包”创建一个模型包模板，创建后可对空模型包进行编辑操作，按照需求添加文件。

支持对Jupyterlab特征工程归档的模型包进行新建模型包操作。

步骤1 单击 

，弹出“新建模型包”对话框。

步骤2 配置“新建模型包”对话框的参数，如**表4-85**所示。

表 4-85 模型包打包配置参数

参数名称	参数描述
模型名称	模型包的名称。
模型版本	模型包的版本。格式为“数值.数值.数值”，其中“数值”为1-2位正整数。
归档包列表	待打包的归档包。 系统自动列出训练平台上当前已经归档的模型包，用户可以通过勾选单个或多个模型包进行打包。 如果需要引入外部模型文件，可不勾选模型列表中的模型，系统将会创建一个模型包模板，用户可以通过对空模型包进行编辑操作，添加模型文件。
模型描述	模型包的描述信息。

步骤3 单击“打包”，系统提示模型打包成功。

----结束

4.8.3 编辑模型包

用户可以使用编辑功能编辑模型包内的文件，上传新文件等。

步骤1 单击模型包“操作”列对应的  。

进入WebIDE模型包编辑界面。

说明

模型包需配置了web ide开发环境才可以进行编辑。如果当前有可用的开发环境，请从模型包记录“开发环境”对应的下拉框内选择可用环境来切换当前环境；如无可用环境，可通过单击“模型管理”界面右上方的“开发环境”，创建一个web ide环境。

步骤2 展开左侧文件管理图标  ，在文件目录中展开模型包同名文件夹，双击待编辑文件在右侧编辑区域进行编辑。

说明

支持简单图形化编辑模型包元数据文件“metadata.json”。在文件目录中单击该文件，右侧编辑区域可编辑代码，也可通过单击代码编辑区域右上角的“  ”打开图形化编辑界面进行编辑，当前该图形化界面支持配置部分元数据内容。

步骤3 右键单击文件目录空白区域，选择“NAIE Upload”，在右侧编辑区域选择上传类型，并从本地选择待上传文件上传。

步骤4 完成模型包编辑后，右键单击文件目录空白区域，选择“NAIE Package”。

说明

完成模型包编辑后必须要执行“NAIE Package”操作，否则编辑无法同步到模型包。

----结束

4.8.4 上架模型包至 AI 市场

步骤1 单击模型包所在操作列的图标。

弹出提交确认提醒。

步骤2 在“确认”弹框内单击“确定”。

系统提示模型包上架到AI市场成功。

----结束

4.8.5 发布推理服务

训练服务支持一键发布在线推理服务。用户基于成熟的模型包，创建推理服务，直接在线调用服务得到推理结果。操作步骤如下。

步骤1 单击模型包“操作”列的，弹出“发布推理服务”对话框，如图4-110所示。

图 4-110 推理服务

步骤2 配置对话框参数如表4-86所示。

表 4-86 创建推理服务参数配置

参数名称	参数描述
模型包名称	发布成推理服务的模型包名称。
版本	推理服务的版本。 版本建议格式为“xx.xx.0”，其中xx为0-99的整数。
是否自动停止	是否开启推理服务自动停止，如果开启，需要设置自动停止的时间，开启了自动停止的推理服务将会在设置时间后停止运行。
计算节点规格	计算节点资源，包括CPU和GPU。 用户可以单击选定计算节点资源，并在“计算节点个数”中配置计算节点资源的个数。
计算节点个数	计算节点的个数。 ● 1代表单节点计算 ● 2代表分布式计算，开发者需要编写相应的调用代码。可使用内置的MoXing分布式训练加速框架进行训练，训练算法需要符合MoXing程序结构。可参考如下文档：https://github.com/huaweicloud/ModelArts-Lab/tree/master/docs/moxing_api_doc
描述	推理服务描述信息。
环境变量	用户可以在训练算法编辑界面中代码目录下predict文件夹中的predict.py文件中编辑推理算法。在创建推理服务的界面中配置环境变量的参数值。 ● 变量名：环境变量的名称 ● 变量值：环境变量的取值 ● 增加：新增环境变量 ●  ：删除环境变量 ●  ：单击可隐藏变量值的真实数据。

步骤3 单击“确定”，发布推理服务。

-  ：发布服务成功，单击图标可以跳转至推理服务的快速验证界面，用户可在此界面上对当前发布的推理服务效果进行验证。
-  ：发布服务失败，可重新发布。

----结束

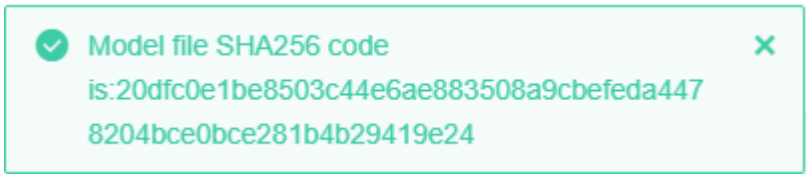
4.8.6 模型包完整性校验

可以对下载的模型包进行完整性校验，判断在下载过程中是否存在篡改和丢包现象。

步骤1 单击模型包操作列的“  ”。

“模型管理”页面右上角展示该模型包的SHA256码，如图4-111所示。

图 4-111 模型包下载前的 SHA256 码



步骤2 单击“”下载模型包，并保存至本地目录。

步骤3 打开本地命令提示符框，输入如下命令生成下载模型包的SHA256码。

```
certutil -hashfile D:\test123-1.0.0.zip SHA256
```

其中“D:\test123-1.0.0.zip”表示模型包的本地下载路径及模型包名，请根据实际情况修改。

得出如下结果：

```
SHA256 的 D:\test123-1.0.0.zip 哈希:
20dfc0e1be8503c44e6ae883508a9cbefeda4478204bce0bce281b4b29419e24
```

步骤4 对比模型包下载前和下载后生成的SHA256码，如果一致，则表示下载过程不存在篡改和丢包。

----结束

4.9 模型验证

4.9.1 模型验证简介

模型验证是基于新的数据集或超参，对训练平台已打包的模型包进行验证，根据验证报告判断当前模型包的优劣。模型验证相关的两个基本概念：

- 验证服务：模型验证服务，编辑和调试模型验证代码。用户可以基于打包成功的模型创建多个验证服务。
- 验证任务：验证服务的实际训练任务。用户加入验证的时候，可以基于指定的模型包，选择不同的数据集、超参和计算资源，创建验证任务。

“模型验证”页面列出了已有的模型验证服务列表信息，如图4-112所示。在该页面，用户可以查看模型验证服务的创建信息，新建、编辑或删除模型验证服务，详情请参见表4-87。

图 4-112 模型验证页面

表 4-87 模型验证页面说明

参数名称	参数说明
名称	验证服务名称。
创建时间	验证服务创建时间。
创建者	创建验证服务的用户。
任务描述	验证服务的描述信息。
	进入验证服务编辑页面，编辑修改验证服务。
	删除验证服务。
FINISHED	当前验证服务创建的最近一次验证任务状态。

4.9.2 创建验证服务

创建验证服务包括三个部分：

- **新建验证服务**：创建验证服务，配置模型类型。具体操作请参考**新建验证服务**。
- **编辑验证代码**：编辑模型验证代码。具体操作请参考**编辑验证代码**。
- **调试验证代码**：调试编辑的验证代码，配置验证代码的调试环境。具体操作请参考**调试验证代码**。

新建验证服务

步骤1 单击“模型验证”页面右上角的“创建”。

弹出“创建验证服务”对话框。

配置“创建验证服务”对话框参数：

- 名称：验证服务名称。只能以字母（A~Z a~z）开头，由字母、数字（0~9）、（_）组成，不能以下划线和中划线结尾，长度范围为[1,26]。
- 描述：验证服务的描述信息。
- 模型类型：从下拉框中选择Tensorflow或Sklearn。Tensorflow和Sklearn分别提供了模版验证代码，用户可以根据实际情况进行勾选。

步骤2 单击“确定”。

进入新建的验证服务详情页面，如**验证服务页面**所示。验证服务页面介绍如**验证服务界面说明**所示。

图 4-113 验证服务页面

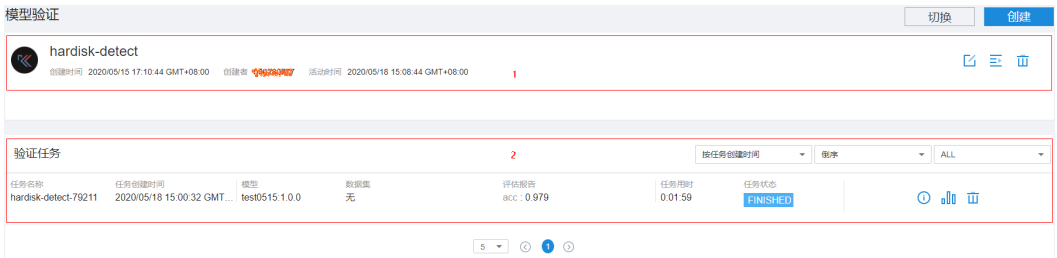


表 4-88 验证服务界面说明

区域	参数名称	参数说明
1（验证服务）	创建时间	验证服务创建时间。
	创建者	创建验证服务的用户。
	活动时间	最近一次验证任务执行的时间。
		进入验证服务编辑界面。
		创建新的验证任务，详细请参考**创建验证任务**。
		删除验证服务。
2（验证任务）		根据状态快速检索验证任务。
		根据任务创建时间、任务名称检索训练任务。 默认按任务创建时间检索。
		按任务创建时间或者任务名称检索训练任务，检索结果按正序或者倒序排列展示。 默认按倒序排序。
	任务名称	验证任务的名称。
	任务创建时间	验证任务创建的时间。
	模型	创建验证任务时选择的模型。
	数据集	创建验证任务时设置的验证数据实例。
	评估报告	验证任务加入验证后，生成的评估报告。
	任务用时	验证任务加入验证的耗时。
	任务状态	验证任务的状态。
		查看验证任务的运行报告，包括系统日志、运行日志和运行图。
		查看验证任务的验证报告。
		删除验证任务。

----结束

编辑验证代码

进入验证代码编辑页面的方式有两种：

- 在模型验证页面，单击验证服务的  ，进入验证代码编辑界面。
- 在验证服务页面，单击右上角  ，进入验证代码编辑界面。

验证代码编辑界面与模型训练中代码编辑界面相似，也分为代码编辑菜单栏、任务执行区域，代码编辑区、代码目录。此处不再详细展开介绍，详情请参见**编辑训练代码（简易编辑器）**。

用户可以在验证代码编辑界面中编辑代码，按下“Ctrl+S”保存当前代码。

调试验证代码

步骤1 单击“代码目录”中的  ，新建与*.py文件对应的*.ipynb文件。

步骤2 单击“Notebook配置”，弹出Notebook配置对话框。

如果有已经创建好的Notebook环境，直接选中“运行中”的环境，单击“保存”即可。否则需要重新创建Notebook开发环境，操作步骤如下：

1. 从AI引擎下拉框中选择AI引擎，单击“创建Notebook环境”。
2. 从调试资源下拉框中选择GPU、CPU调试资源。
3. 待环境状态为“运行中”时，选中该环境。
4. 单击“保存”。

步骤3 单击“*.ipynb”文件进入算法调试界面。

步骤4 依次单击“Kernel > Change kernel”，选择AI引擎。

步骤5 在输入框中配置算法，单击 

调试算法。

如果调试界面的Cell区没有报出异常，说明算法运行正常，用户可以基于该算法对验证数据集进行训练。

----结束

4.9.3 创建验证任务

验证任务主要是对指定的模型包，基于调试好的验证代码，设置新的数据集、超参和计算资源，执行验证任务，验证该模型包的优劣。

创建验证任务的方法有两种：

- 在验证代码编辑页面，单击右上角的“验证”，弹出“验证配置”对话框。配置“验证配置”对话框参数，创建验证任务。
- 在验证服务页面，单击右上角的  ，弹出“验证配置”对话框。配置“验证配置”对话框参数，创建验证任务。

此处以在验证代码编辑页面创建验证服务为例，配置方法如下。

步骤1 在验证代码编辑页面，单击右上角的“验证”。

弹出“验证配置”对话框。参数配置如表4-89所示。

表 4-89 验证配置参数说明

参数名称	参数说明
验证模型	待验证的模型包。下拉框中列出了系统当前已打包的模型包。
验证数据集	按行展示每个数据集超参内容。每行分别展示数据集超参名称、数据集名称、经过特征处理后生成的数据实例名称。标签列如果通过“运行超参”设置，此处可置为空。
参数配置	重新配置验证任务时的参数。用户可以通过 + 新增参数。
AI引擎	AI引擎及AI引擎的版本。
计算节点规格	系统提供的计算节点资源。

步骤2 单击“创建”，创建验证任务。

用户可以单击验证代码编辑页面右上角 

，或者直接在验证服务页面，查看验证任务执行情况：

-  ：验证任务执行过程中，动态查看验证任务的系统日志、运行日志和运行图。
-  ：验证任务结束后，查看验证任务的验证报告，当前评估报告仅支持数值类型。

在验证任务执行过程中，用户可以单击  停止任务。

----结束

4.10 云端推理框架

4.10.1 推理服务

云端推理框架提供模型云端运行框架环境，用户可以在线验证模型推理效果，无须从零准备计算资源、搭建推理框架，只需将模型包加载到云端推理框架，一键发布成云端Web Service推理服务，帮助用户高效低成本完成模型验证。

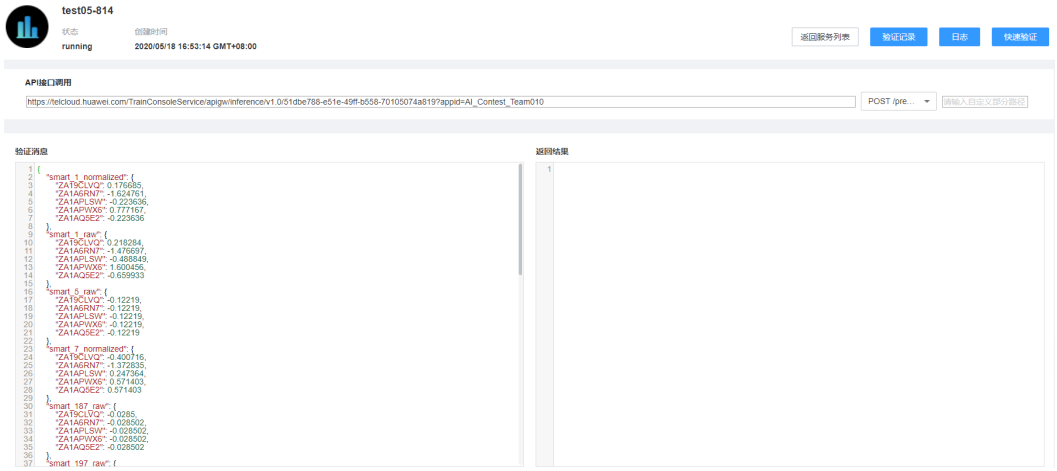
其中，“推理服务”主界面默认展示所有推理服务，用户可查看推理服务详情，并对推理服务进行一系列操作。

步骤1 在模型管理界面，单击模型包所在行，对应操作列的  。

进入推理服务验证页面，如图4-114所示。支持用户在界面上直接验证模型推理效果。

页面中显示了推理服务的API接口信息。系统默认支持“POST / ”接口，并且支持在模型包中自定义REST接口，通过在线推理发布成REST服务。

图 4-114 推理验证



步骤2 在页面左侧“验证消息”区域中输入json格式的数据，单击“快速验证”。

右侧“返回结果”区域框会给出在线推理结果。

步骤3 在菜单栏中，选择“推理服务”。

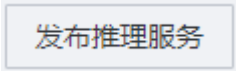
进入推理服务主页面，页面以图表的形式展示所有推理服务，如图4-115所示。

图 4-115 推理服务

界面说明如表4-90所示。

表 4-90 推理服务界面说明

区域	参数	参数说明
1	请输入名称	支持通过名称快速检索推理服务。
	全部状态	支持通过推理服务的状态快速筛选相同状态下的推理服务。 状态包括：运行中、部署中、已停止、告警、部署失败、后台异常。

区域	参数	参数说明
		用于刷新推理服务界面内容。
		使用“模型仓库”中的模型包创建新的推理服务。
2		查看推理服务详情信息，包括：CPU/内存/GPU当前使用量、模型包详情、历史更新记录和事件详情。
		推理服务的日志。支持通过“自定义时间段”筛选日志。
		支持复制推理服务提供的API接口地址。
		进入快速验证界面，支持用户在界面上直接验证模型推理效果。
		将推理服务授权给其他用户使用。 服务发布者将推理服务授权给他人后，授权界面会生成“访问地址”，被授权用户可凭借自己的账户Token和“访问地址”调用推理服务的API接口。
		停止运行推理服务。
		修改推理服务的配置信息，包括是否开启自动停止、版本、计算节点规格、计算节点个数、分流、环境变量。
		删除推理服务。

说明

推理服务主页面快速入口：

训练平台首页左上角品牌Logo旁Home图标的标识内容，单击展开下拉选项，选择“推理服务”。

----结束

4.10.2 模型仓库

在菜单栏中，选择“模型仓库”。

进入“模型仓库”界面。界面以列表的形式，展示了当前租户下面已成功创建推理服务的模型包列表和模型包详细信息，如图4-116所示。

图 4-116 模型仓库

界面说明如表4-91所示。

表 4-91 模型仓库界面说明

区域	参数	参数说明
1	请输入关键词	支持通过模型包名称快速检索模型包。
	导入模型包	支持用户通过本地上传或者AI市场导入的方式，导入模型包。
2	模型包名称	模型包的名称。
	版本	模型包生成时的版本。
	模型类型	模型的AI算法框架类型。
	运行环境	AI算法框架匹配的Python语言版本。
	创建时间	模型包生成的时间。
	来源	模型包的来源。包括训练平台、本地上传和AI市场导入三种来源。
	状态	模型包的状态。
	操作	可以对模型包执行下述操作： • ：查看模型包信息，包括名称、版本、描述、基本信息、运行依赖。 • ：将模型包发布成推理服务。 • ：删除模型包。

4.10.3 模板管理

云端推理框架新增模板能力，用户在云端推理框架发布推理服务时，可以使用系统预置的模板，将模型包发布成推理服务。

背景信息

在训练服务“模型管理”界面发布的推理服务，仅封装了Tensorflow类型的模型。对模型包格式上限制导致定制会比较多。或者使用特殊环境的Case难以实现，比如：KPI

异常检测服务使用了很多Python的框架，且需要自定义启动的方式；有些Case还需要使用Java、Tomcat。

包括如下缺点：

1. 对模型包格式有约束。虽然云端推理框架对训练服务发布的推理服务，进行了适配和封装，例如：预置若干必须的文件。还是对开发者增加了隐含约束，比如：流量预测服务曾遇到模型被覆盖的问题。
2. 对入口文件“custom_service.py”的实现方式有约束，必须实现特定的接口，如：TensorflowService。如果推理服务不使用Tensorflow引擎，实现起来效果不理想。
3. 仅支持提供一个推理服务调用接口，无法满足某些Case的需求，比如：KPI异常检测。

模板优势

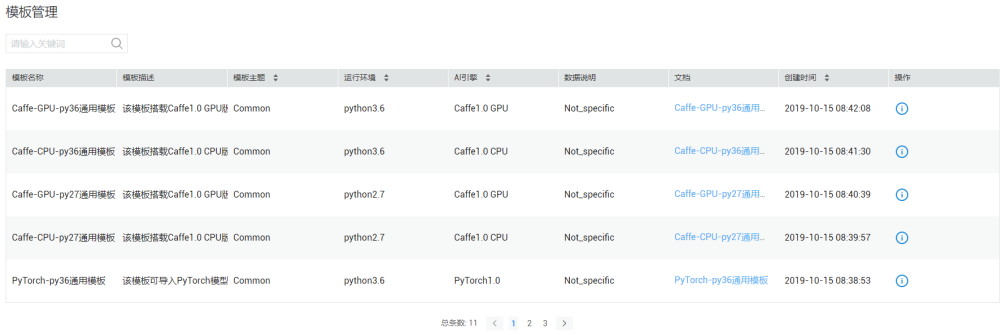
使用云端推理框架的“模板管理”具备如下优势：

相对于仅能使用固定类型的模型类型TensorFlow，模板部署模型包的方式仅可以满足定制化的需求。比如：使用Java的Case；KPI异常检测Case定制启动的命令或提供多个推理服务调用接口。

模板管理界面说明

“模板管理”界面以列表的形式，展示了当前租户下面已创建成功的模板列表和模板详细信息，如图4-117所示。

图 4-117 模板管理



界面说明如表4-92所示。

表 4-92 模板管理界面说明

参数	参数说明
请输入关键词	搜索模板名称关键字，快速查找模板。
模板名称	模板的名称。
模板描述	模板的描述信息。
模板主题	模板的主题。支持按照首字母进行顺序排列或倒叙排列。

参数	参数说明
运行环境	AI算法运行的环境。支持按照首字母进行顺序排列或倒叙排列。
AI引擎	AI算法框架。
数据说明	数据说明信息。
文档	跟模板相关的文档名称，单击文档名称支持跳转至文档内容界面。
创建时间	模板创建的时间。
操作	可以对模板执行下述操作：  ：查看模板配置信息。

4.11 修订记录

发布日期	修订记录
2020-09-30	数据集详情界面优化，更新**新建数据集和导入数据**。 **基于Jupyterlab的自动机器学习**章节，针对AutoML自动机器学习，输出场景化资料。 模型管理界面优化，更新**模型管理**。 删除模型管理界面的云端推理入口，更新**云端推理框架**。
2020-08-17	根据最新的训练平台，更新“**训练服务简介**”章节描述。 **新建数据集和导入数据**章节“支持超大文件（10G）上传”功能增强。 模型训练任务界面优化，对应刷新**模型训练**截图及界面参数描述。 模型验证任务界面优化，对应刷新**模型验证**截图及界面参数描述。
2020-07-16	新增“**学件**”章节。 **数据集简介**章节新增“DatasetService数据集”介绍。 **新建数据集和导入数据**章节新增“支持超大文件（10G）上传”操作指导。 训练任务页面优化，对应刷新**模型训练**截图。 推理服务API接口优化，对应修改**推理服务**。
2020-06-16	模型训练新增MindSpore样例体验，对应刷新**模型训练**。 新增Tensorboard管理，对应刷新**模型训练**。

发布日期	修订记录
2020-05-18	变更点如下： • Jupyterlab特征工程编辑界面菜单调整、新增时序数据算子、新增Box-Cox变换、优化模型训练、特征迁移增加迁移评估等，对应刷新**JupyterLab开发平台**。 • 模型训练新增创建联邦学习工程及其服务，对应新增**创建联邦学习工程**。 • 模型包支持对Jupyterlab特征工程归档的模型创建模型包、支持对特定模型包新建联邦学习实例、支持对已发布推理服务的模型包更新发布推理服务，对应刷新**模型管理**。
2020-04-16	变更点如下： • 训练服务首页项目列表“开发环境”列优化，对应刷新**训练服务首页简介**。 • Jupyterlab特征工程功能变更，对应刷新**JupyterLab开发平台**。 • 模型训练功能优化，对应刷新**模型训练**。 • 模型管理新增模型包完整性校验，对应新增**模型包完整性校验**。
2020-03-30	JupyterLab开发平台界面和功能优化，对应刷新“**JupyterLab开发平台**”章节全量内容。 模型训练服务的“模型训练”菜单界面优化，对应刷新“**模型训练**”章节全量内容。 模型训练服务的“模型管理”页面增加推理服务入口，对应刷新“**发布推理服务**”章节内容。
2019-12-30	新增如下章节： • 订购训练服务 • **训练服务首页简介** • **JupyterLab开发平台** • **编辑训练代码（WebIDE）** • **创建模型包** • **编辑模型包** • **上架模型包至AI市场** • **发布推理服务** • **云端推理框架**
2019-10-30	特征工程编辑界面的菜单优化，对应“特征工程”章节内容调整和优化。 新增如下章节： • **Notebook开发** • **创建超参优化服务** • **创建Tensorboard**
2019-04-30	第一次正式发布。

5 学件开发指南

5.1 学件能力简介

背景

网络AI特性开发业务活动中，对很多运维场景有共性需求，比如异常检测、故障定位、故障预防预测等。以KPI异常检测场景为例，存在如下共性需求：

- 运营商和企业客户对于KPI实时监控，快速定位故障有共性需求。
- 运营商网络中存在海量KPI。例如：路由器有70000+KPI，其中丢包和统计类有4000+KPI。
- DCN对接口/设备KPI、光链路、VM/应用均有异常检测需求。

针对KPI异常检测场景，缺乏公共算法能力积累，异常检测模型开发效率低，成本高。存在如下问题：

- 产品对异常检测需求持续增加，单个异常检测模型开发周期约6个月，无法快速生成模型。
- 同时需要投入1至2名算法专家进行数据清洗、特征分析、模型选择和验证等工作，模型开发成本高。

学件概念

学件能力，支持部分重用他人结果，不必“从头开始”。

学件（Learnware）= 模型（model）+ 规约（specification）

其中，规约需要能够给出模型的合适刻画，模型需要满足如下条件：

- 可重用：不用用户之间可分享模型，不需要分享数据，避免了数据隐私和数据保护。
- 可演进：学件本身需要可演进，能适应环境，可增量学习
- 可了解：规约需要给出模型适应场景。

学件还具备如下特点和优势：

- 可不依赖数据：通过数据训练好的模型提供出去。把参数、网络结构等内容提供给出去，不提供数据，解决数据安全问题。

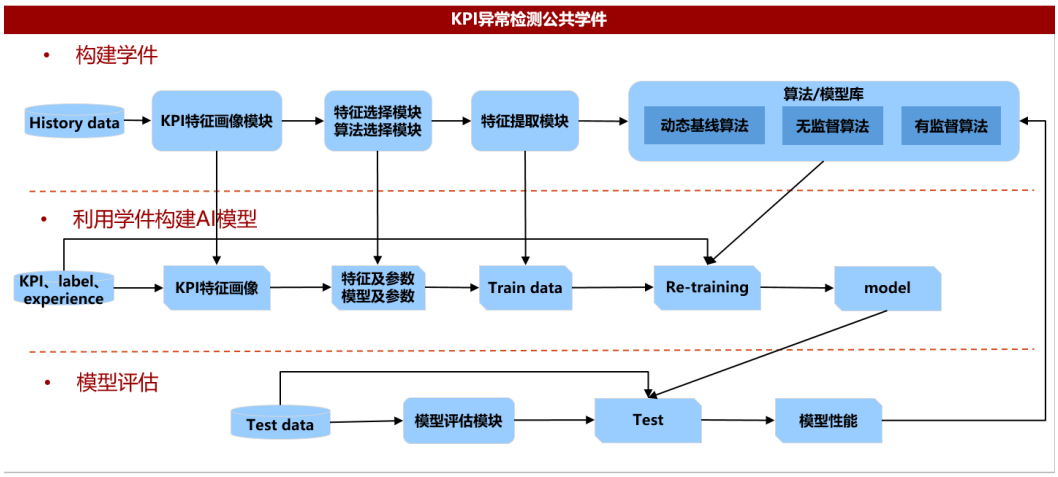
- 可不依赖专家：具备基础模型，在约定的模型适应场景中可部分重用。

KPI 异常检测公共学件

异常检测学件服务，通过数据特征画像识别数据类型，自动推荐训练算法与特征，采用无监督、有监督和动态基线等进行联合检测，通过专家经验对训练与检测进行调优，得到最终检测结果。模型训练完成后，可以将特征画像的结果、特征和参数、模型和参数都保留下来。后面仅需要使用新的数据，重训练模型，不用再重新做特征分析和模型分析。目前，学件已经集成了几十维到上百维不同种类的特征库，源于历史各类Case和通用KPI异常检测的算法库。后面会不断丰富特征库和算法库。

KPI异常检测公共学件能力，如图5-1所示。

图 5-1 KPI 异常检测公共学件



KPI异常检测公共学件的功能，如表5-1所示。

表 5-1 公共学件的功能模块

功能模块	说明
数据接入模块	实现与各类数据源的接口、格式转换等。
数据管理模块	提供源数据、标注样本的存储、导入导出、查询等功能。
数据处理模块	主要实现数据的预处理，包括标签处理、缺失值填充、数据标准化等。
特征工程模块	主要实现对KPI的数据分布特征进行分析，自动选择特征及参数。并提供四大类，80+特征的自动提取。
模型管理模块	主要实现根据KPI的标签、数据分布特征等进行异常检测算法的自动选择、参数设置及模型训练、推理。
数据交互模块	主要支撑公共学件与用户的交互，包括数据管理、数据的可视化展示、专家经验注入等。

5.2 订购模型训练服务

步骤1 在用户终端PC上打开浏览器，在地址栏中输入“**https://console.huaweicloud.com/naie/**”，进入AI市场。

用户首次访问AI市场，会进入“访问授权”界面，单击“授权并继续”即可。

步骤2 单击界面右上角的“登录”，进入登录界面。

步骤3 输入租户名和密码，单击“登录”，进入AI市场。

首次登录后请及时修改密码，并定期修改密码。

步骤4 依次选择“AI服务 > 模型与训练服务 > 模型训练 > 模型训练服务”，进入模型训练服务介绍页面。

步骤5 单击“我要购买”，进入如图5-2所示的界面。

区域：为用户提供服务的华为云Region。请选择“华北-北京四”。

用户可以单击“了解计费详情”，详细了解训练服务提供的资源、规格和相应的价格信息。同时，用户在使用具体资源时，训练服务会在界面给出醒目的计费提示。

图 5-2 订购训练服务

步骤6 单击“立即使用”，服务订购完成。

----结束

5.3 访问模型训练服务

步骤1 在用户终端PC上打开浏览器，在地址栏中输入“**https://console.huaweicloud.com/naie/**”，进入AI市场。

步骤2 单击界面右上角的“登录”，进入登录界面。

步骤3 选择“IAM用户登录”方式，输入租户名、用户名和密码。

用户也可以直接通过账号登录。首次登录后请及时修改密码，并定期修改密码。

步骤4 单击“登录”，进入AI市场。

步骤5 依次选择“AI服务 > 模型与训练服务 > 模型训练 > 模型训练服务”，进入模型训练服务介绍页面。

步骤6 单击“进入服务”，进入模型训练服务页面。

----结束

5.4 KPI 异常检测学件服务

5.4.1 创建项目

KPI异常检测学件服务，封装在模型训练服务的“KPI异常检测”模板中。可通过创建“KPI异常检测”模板项目，体验KPI异常检测学件服务。

步骤1 在训练平台首页，单击“KPI异常检测”模板中的“使用模板创建”。

弹出“创建项目”对话框，如图5-3所示。

图 5-3 创建项目

步骤2 配置“创建项目”对话框参数，如表5-2所示。

表 5-2 参数说明

参数名称	参数说明
名称	项目的名称。 名称只能以字母（A~Z a~z）开头，由字母、数字（0~9）、下划线（_）、中划线（-）组成，不能以下划线和中划线结尾，且长度为[2-20]个字符。
描述	对项目的简要描述。 字数不能超过500。
开发环境	项目创建后，会创建对应规格的JupyterLab平台。JupyterLab平台封装了学件相关的能力，用户可以进行特征画像、特征选择和参数设置、算法选择和参数设置、模型训练和评估。
是否公开	项目是否可以被所属用户组的其他用户访问： • 是 • 否
公开至组	仅当“是否公开”设置为“是”，才会展示“公开至组”。 默认展示当前用户所属的所有用户组，如果勾选用户所属的用户组，则被勾选用户组下的所有用户均可以查看当前项目。
图标	项目图标。 支持用户本地上传。

步骤3 单击“创建”，完成KPI异常检测模板项目的创建。

----结束

5.4.2 数据集

学件项目中预置了3份样例数据，《 学件开发指南 》手册选择其中一份样例，讲解学件的操作流程。

如果用户需要使用自己的数据，可以参考**新建数据集和导入数据**，创建新的数据集，并导入数据。

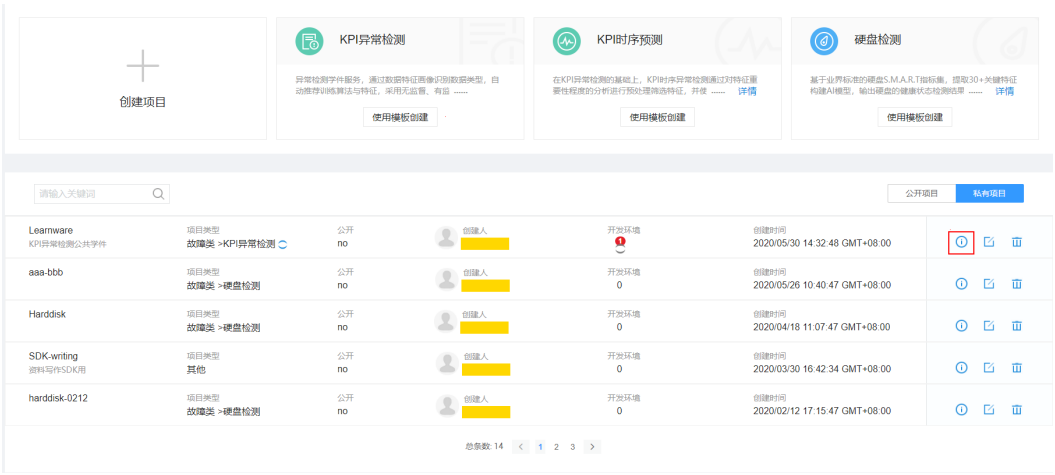
导入数据要求

- 建议训练数据和测试数据分成两个实例，方便算法查找训练或测试数据的位置。
- 训练数据可以是带标签或者不带标签的数据，测试数据一定是带标签的数据，方便评估模型执行效果。

查看学件项目预置的样例数据

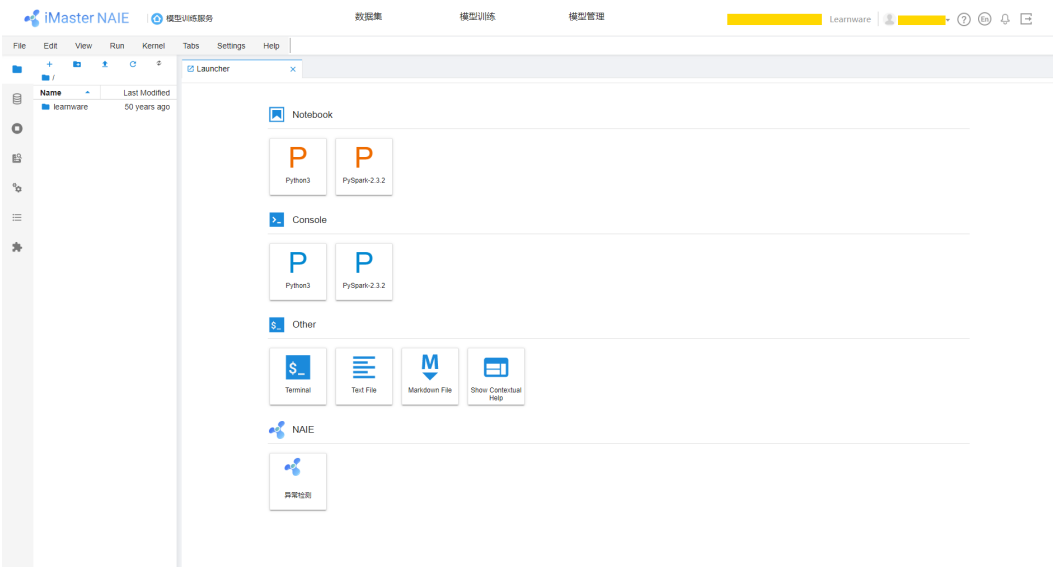
步骤1 等待学件项目创建完成后，在训练平台首页的项目列表中，找到创建完成的学件项目。单击项目所在行的 图标。

图 5-4 学件项目



步骤2 进入项目编辑界面，如图5-5所示。

图 5-5 项目编辑界面



步骤3 在菜单栏中，单击“数据集”，进入“数据集”界面，如图5-6所示。

查看学件项目中预置的三类样例数据（Unlabeled、Gpr和Labeled）。使用每类样例数据体验学件能力，会对应到不同的算法，训练生成不同的模型。

图 5-6 数据集界面

----结束

新建数据集和导入数据

步骤1 在数据集菜单页面，单击界面左上角的 图标。

弹出“导入数据”对话框，如图5-7所示。

图 5-7 导入数据

配置“导入数据”对话框参数，具体参见表5-3。

表 5-3 导入数据参数说明

参数名称	参数说明
数据集	输入自定义名称。当“导入数据”，单击“创建”后，自动完成新数据集的创建。
数据类别	导入数据的类别。
实例名称	本次导入数据的名称。
实例别名	本次导入数据的别名。

参数名称	参数说明
数据来源	数据上传的途径。 包含如下方式： ● 本地上传：从用户本地上传数据。 ● 数据目录：导入用户在数据服务的数据集服务中订阅的数据。 ● 样例数据：训练平台环境中预置的用户体验数据。包括鸢尾花原始测试集、鸢尾花训练集、鸢尾花测试集、KPI 15分钟数据集、KPI 60分钟数据集、KPI异常检测数据集。 其中鸢尾花原始测试集、KPI 15分钟数据集和KPI 60分钟数据集中包括空值，用户可以通过特征工程进行数据修复，剔除空值。
本地上传-文件大小限制为80M，文本支持csv和txt	数据来源选择“本地上传”时可见，表示数据文件所在的用户本地路径。 为避免后续处理数据时出错，请按要求上传csv和txt格式的数据文件。
数据目录-请选择数据集	数据来源选择“数据目录”时可见。 选择数据集服务中订阅的数据。 ● 订阅：单击“订阅”图标，自动跳转至数据湖的数据集服务界面，可以查询并订阅数据。 ● 刷新：刷新展示数据集服务订阅的数据列表。 ● 数据名称：数据集服务订阅的数据名称。 ● 申请状态：数据集服务订阅数据的申请状态。 ● 审批人：数据集服务订阅数据的审批责任人。 ● 数据来源：数据集服务订阅数据的来源。 **说明** 在订阅数据目录数据前，需要用户阅读《使用须知》，并签署同意遵守使用敏感数据项目条款或条件约束。
分隔符	用户根据导入数据文件的格式进行选择，用于系统识别数据字段。 当前支持“,”、“;”和“|”三种分隔符。
文件编码	数据文件的编码格式。 当前支持UTF-8、GBK和GB2312三种格式。
标题行	数据是否包含标题行，用户根据导入数据文件的格式进行选择。 包含如下选项： ● 有标题行 ● 无标题行

步骤2 单击“创建”，导入数据文件。

如果导入数据所在的“状态”列显示“导入成功”，说明数据导入成功。

步骤3 (可选)分析数据。

1. 单击状态为“导入成功”的数据，对应“操作”列的图标，进入数据详情界面。
2. 单击“分析数据”。

   分析完成后展示该数据实例的数据详情，包括：字段名称、字段类型、数据分布、有效值、空值、异常值、最大值、最小值、均值、方差、分位数等。
3. 单击界面右上角的“预览数据”，查看具体的数据内容。

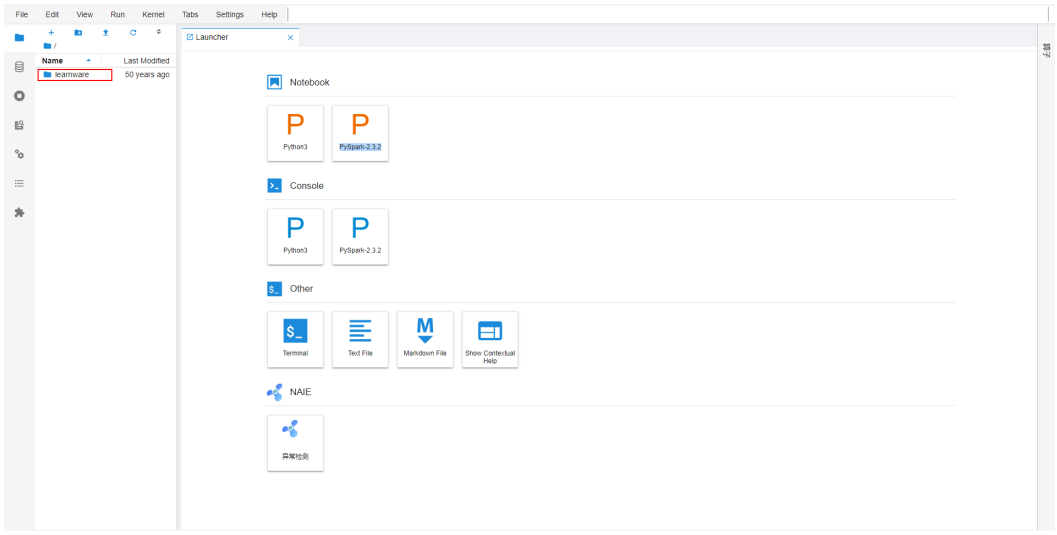
----结束

5.4.3 模型训练

5.4.3.1 导入 SDK

步骤1 在学件项目中，单击菜单栏中的“模型训练”，进入JupyterLab平台界面，如图5-8所示。

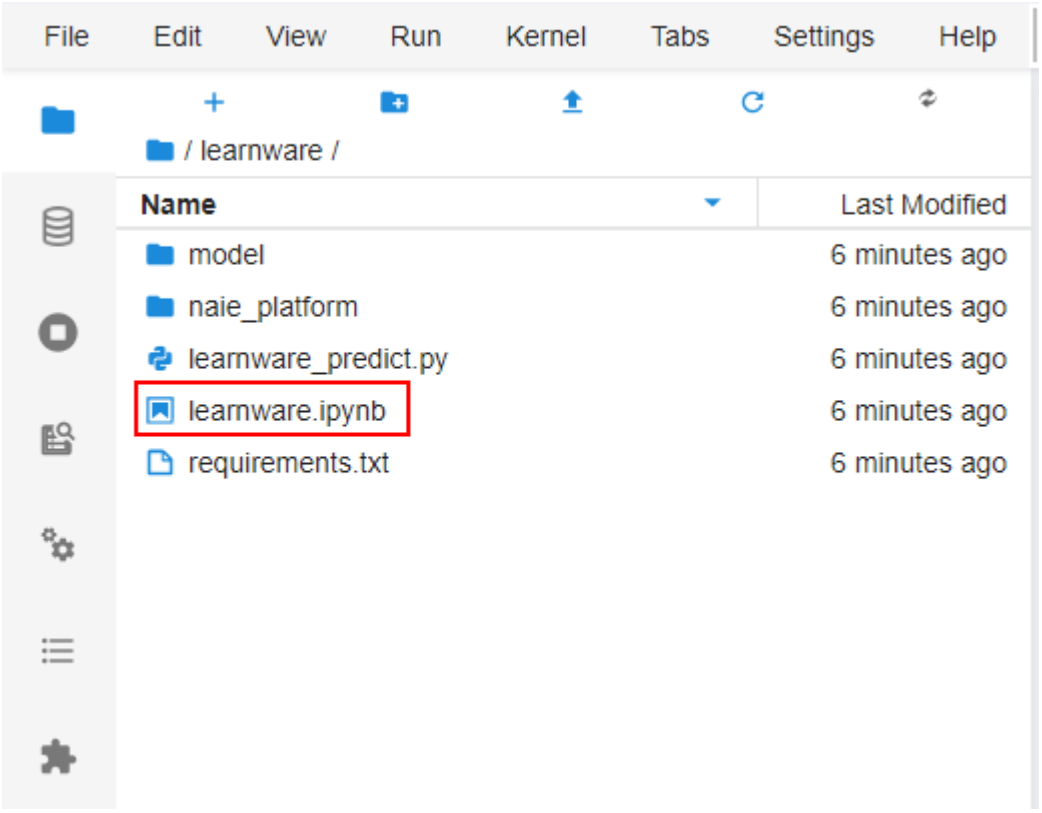
图 5-8 项目编辑界面



步骤2 双击左侧目录中的项目名称“learnware”。

进入“learnware”目录中，如图5-9所示。

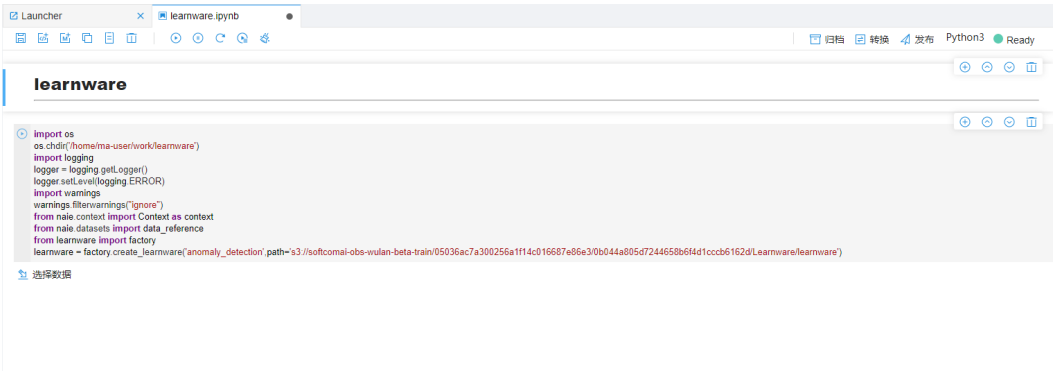
图 5-9 学件项目存储目录



步骤3 双击左侧的“learnware.ipynb”文件。

打开学件项目的jupyterlab环境编辑界面，如图5-10所示。

图 5-10 jupyterlab 环境编辑界面



步骤4 单击第一个代码框左侧的图标，导入算法依赖的训练平台SDK。

----结束

5.4.3.2 选择数据

模型训练前，需要选择训练数据和测试数据。建议训练数据和测试数据分成两个实例，方便算法查找训练或测试数据的位置。

步骤1 单击第一个代码框下方的“选择数据”。

弹出“选择数据”代码框，如图5-11所示。界面对训练集、验证集和测试集的概念做出了详细的注释。

图 5-11 选择数据

选择数据代码框右侧的参数说明，如表5-4所示。

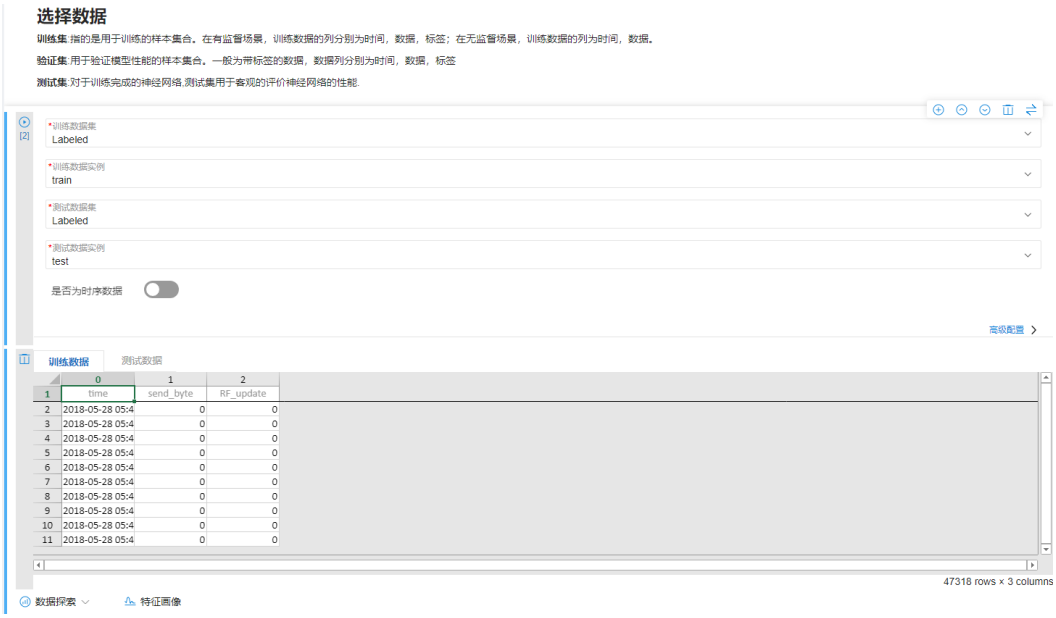
表 5-4 选择数据

参数	参数说明
训练数据集	从下拉框中选择数据集“Labeled”。
训练数据实例	从下拉框中选择训练数据“train”。
测试数据集	从下拉框中选择数据集“Labeled”。
测试数据实例	从下拉框中选择训练数据“test”。
是否为时序数据	请保持关闭。 如果开启，则需要配置如下参数： ● 时间列：输入时间列名称。 ● 时间格式：指定时间字段的时间格式。 ● ID列：数据的标识列。 ● 是否检测周期与平稳性：开启开关会检测时序数据的周期，或判断指定的周期是否为时序数据的周期，以及检测时序数据是否平稳。 如果开启此开关，运行时间会较长，默认关闭此开关。
数据引用变量名	当特征工程需要选择多份数据时，使用此参数给每份选定的数据命名，以免产生冲突。 均保持默认值即可。

步骤2 单击“选择数据”代码框左侧的图标。运行代码，绑定训练和测试数据实例。

运行成功后，可以查看训练数据和测试数据，如图5-12所示。

图 5-12 查看数据



----结束

5.4.3.3 特征画像

特征画像的作用，就是对数据进行分析，把其中一些基本特征提取出来，如：周期性、离散度、时序规律、最值、采样频率等，计算KPI曲线特点（包括周期性、趋势性、噪声、离散性、随机性等）。根据计算的曲线特点，判断KPI的大类别（毛刺型、阶梯型、周期型、离散型、稀疏型、多模态型等）。这些类别，对应到后面的特征选择、算法推荐，会有不同的策略，有效提升模型的构建效率。

步骤1 单击“选择数据”左下方的“特征画像”。

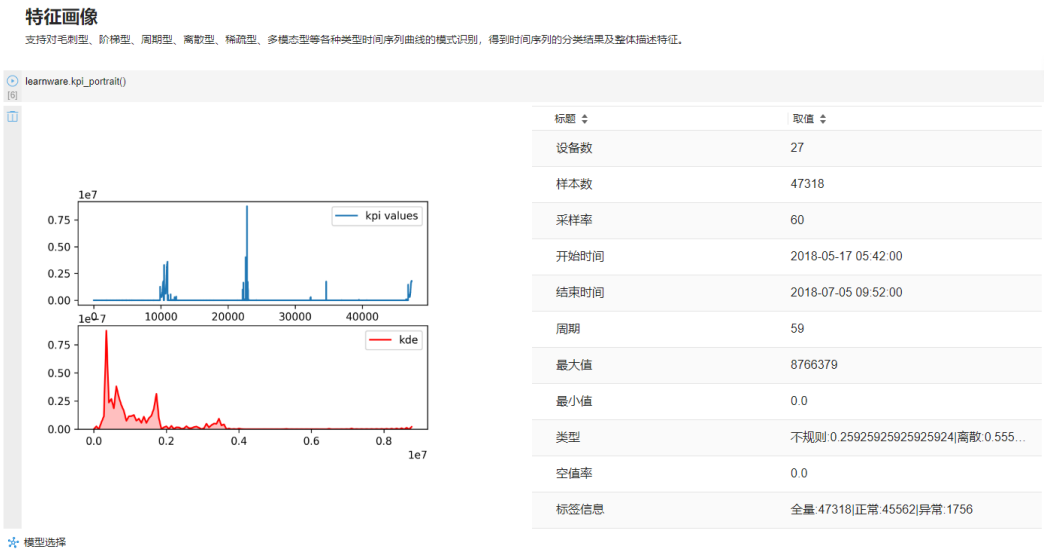
新增“特征画像”内容，如图5-13所示。

图 5-13 特征画像

步骤2 单击“特征画像”代码框左侧的图标，运行代码。

运行结果如图5-14所示。通过左侧两个图可以直观的看一下原始数据和数据的密度分布图。

图 5-14 特征画像结果



右侧的参数说明，如表5-5所示。

表 5-5 特征画像参数说明

参数	说明
设备数	需要检测的KPI对象的数量，如设备或端口的数目。
样本数	训练数据总的样本数。
采样率	采样频率，单位为秒。60的含义为每60秒采样一次。
开始时间	采样的时间跨度。
结束时间	
周期	是否有周期的特性，给出评估的值。
最大值	KPI的最大值。
最小值	KPI的最小值。
类型	KPI类型的计算。
空值率	有没有缺失值。取值为“0”说明，没有缺失值。
标签信息	统计标签的样本数量。

----结束

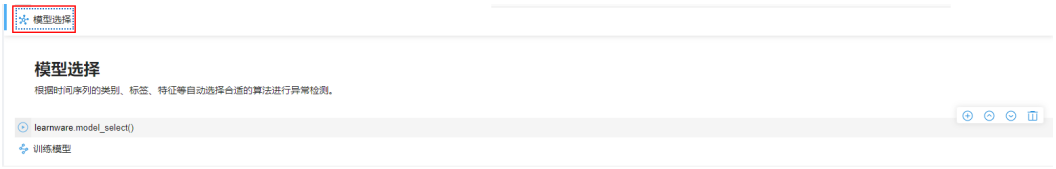
5.4.3.4 模型选择

目前，学件已经集成了几十维到上百维不同种类的特征库，源于历史各类Case和通用KPI异常检测的算法库。通过数据的特征画像，可以实现自动化的特征推荐和算法推荐。

步骤1 单击“特征画像”左下方的“模型选择”。

新增“模型选择”内容，如图5-15所示。

图 5-15 模型选择



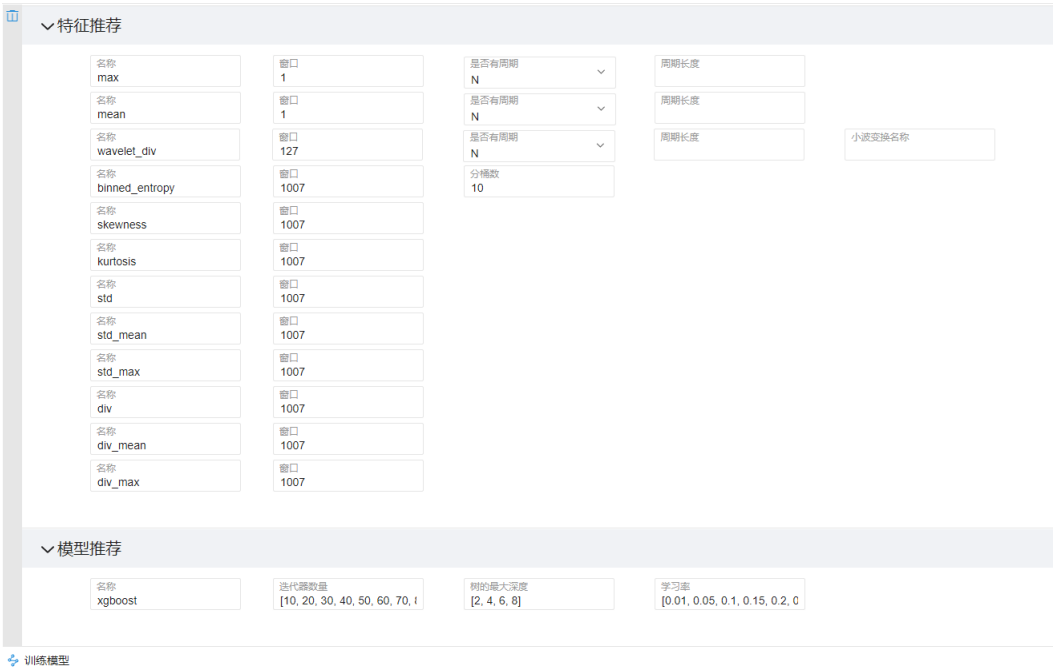
步骤2 单击“模型选择”代码框左侧的 图标，运行代码。

学件推荐的特征列表和模型如图5-16所示。

- 特征推荐：学件推荐的特征，除了一些通用的特征（最值、均值等），还有一部分是专门为类似KPI做的异常检测效果比较好的特征。通常采用滑窗的方式做异常检测。目前所有窗口的长度，是根据数据的周期性、样本数、周期的个数等数据特点推荐的。窗口的长度均可以修改，如果用户对算法比较了解，对当前KPI比较熟悉，可以修改为用户认为更合适的值。
- 模型推荐：前面选择的数据是有标签的数据，推荐算法xgboost是有监督的算法。模型推荐里面增加了超参搜索的功能，有给出参数取值的推荐区间。用户也可以根据实际情况修改。

  如果推荐的是无监督的异常检测算法，可能会同时推荐几个算法。那模型训练的时候，针对不同的算法，会分别进行模型训练，得到不同的模型，通过集成学习投票法策略，推荐得到更符合且更准确的异常检测模型。

图 5-16 特征推荐和模型推荐



----结束

5.4.3.5 训练模型

特征和算法确定后，可以开始训练模型。

训练模型

步骤1 单击“模型选择”左下方的“训练模型”。

新增“训练模型”内容，如图5-17所示。

图 5-17 训练模型

步骤2 单击“训练模型”代码框左侧的 图标，进行模型训练。

训练模型的评估效果，如图5-18所示。

第一列内容的含义如下所示：

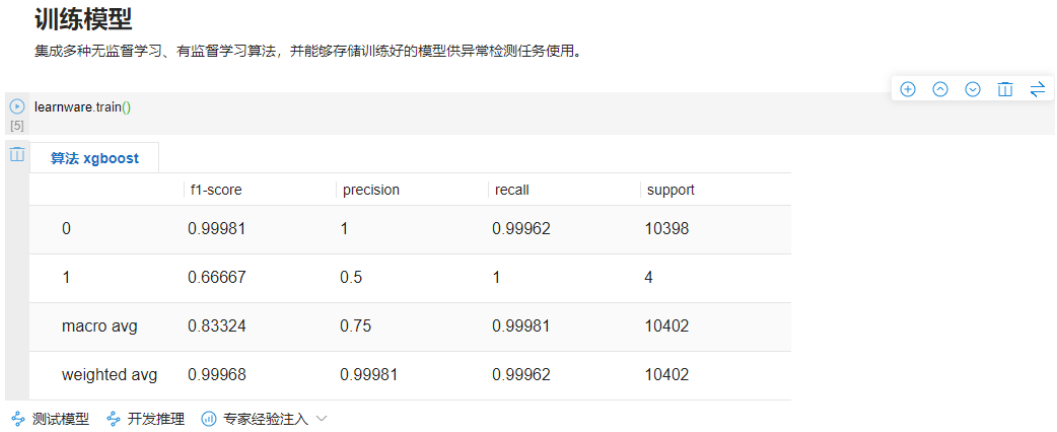
- 0：标注为0的所有样本。可以理解为标签。
- 1：标注为1的所有样本。可以理解为标签。
- macro average：所有标签结果的平均值。
- weighted average：所有标签结果的加权平均值。

模型优劣的评价指标：

- f1-score：F1分数同时考虑精确率和召回率，让两者同时达到最高，取得平衡。
- precision：精确率，又被称为查准率，是针对预测结果而言的。含义为在被预测为正的样本中实际为正样本的概率。
- recall：召回率，又被称为查全率，是针对原样本而言的。含义为在实际为正的样本中被预测为正样本的概率。
- support：每类标签出现的次数。

模型训练完成后，可以查看归档的模型文件，如模型训练目录说明所示。

图 5-18 训练模型结果



----结束

模型训练目录说明

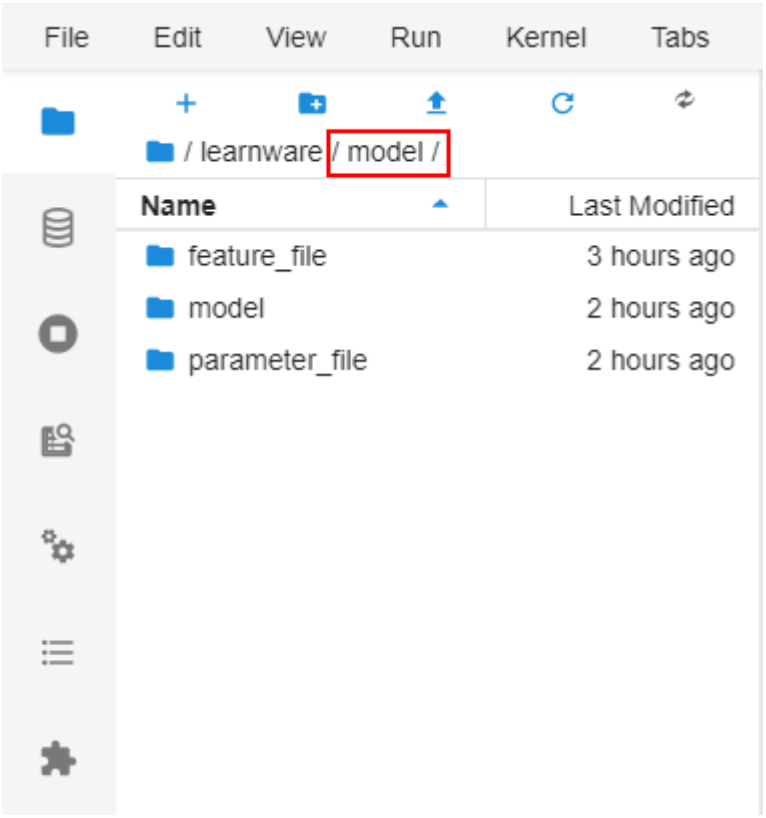
模型训练完成后，训练好的模型和相关内容，都保存在如图5-19所示的model目录中。将model目录导出，使用新数据，直接利用已有的特征和参数、算法和参数，就可以实现模型重训练。

model目录的上级目录“learnware”是用户创建的学件项目名称。

model目录的子目录含义如下所示：

- feature_file：存放推荐的特征配置列表文件和KPI特征画像文件。
- model：存放训练好的模型。
- parameter_file：存放模型推荐的算法和参数配置文件。

图 5-19 model 目录



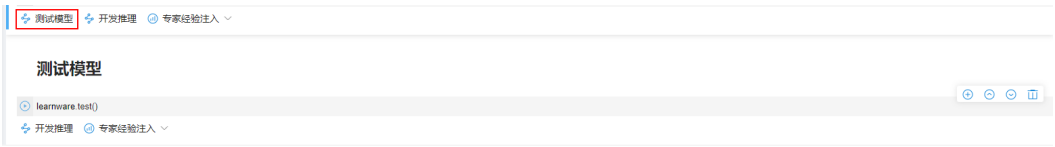
5.4.3.6 测试模型

用测试数据测试模型的泛化能力。训练数据可以是带标签或者不带标签的数据，测试数据一定是带标签的数据，方便评估模型执行效果。

步骤1 单击“训练模型”左下方的“测试模型”。

新增“测试模型”内容，如图5-20所示。

图 5-20 测试模型



步骤2 单击“测试模型”代码框左侧的 图标，进行模型评估。

模型的评估效果，如图5-21所示。

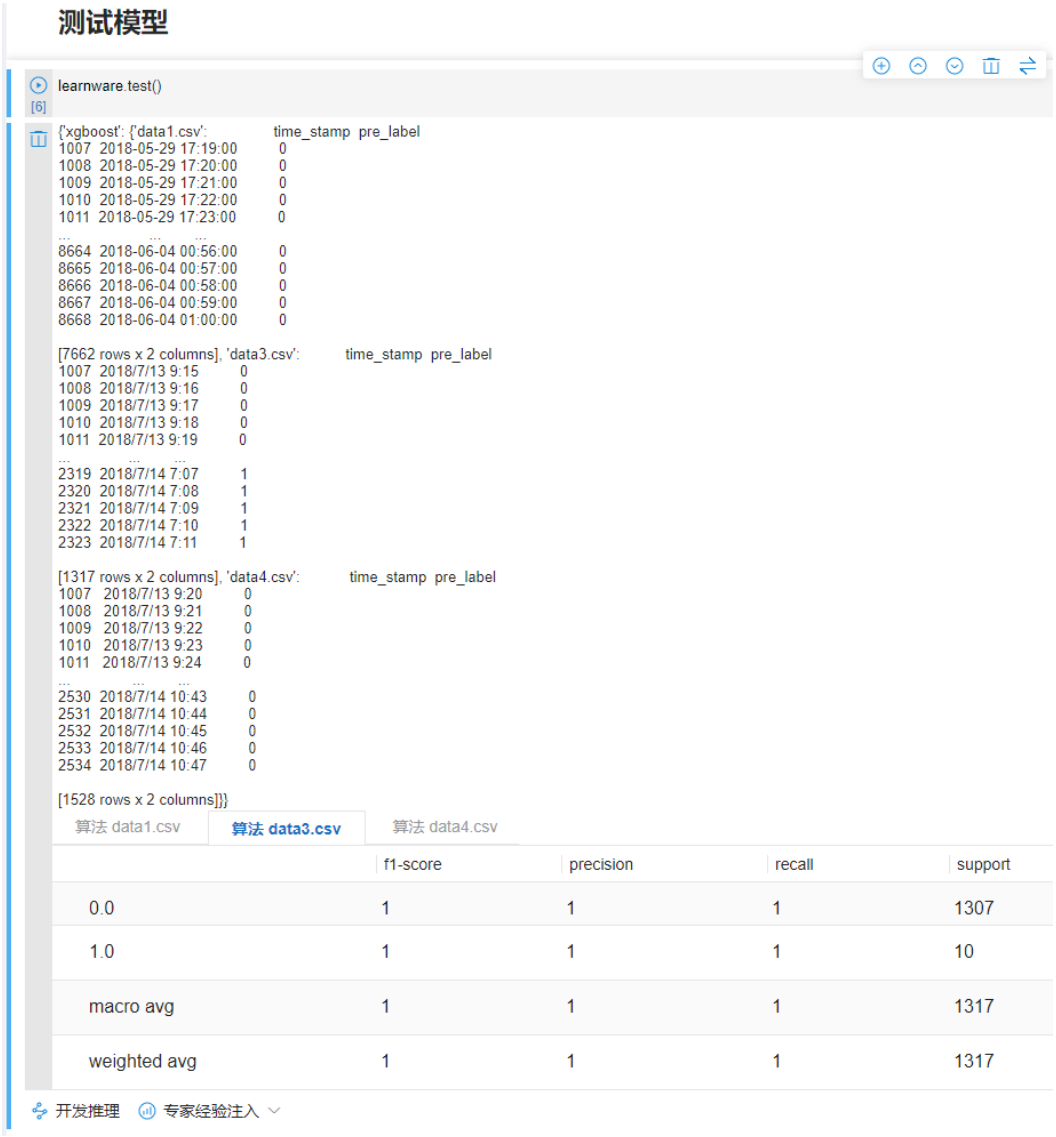
第一列内容的含义如下所示：

- 0.0：标注为0的所有样本。可以理解为标签。
- 1.0：标注为1的所有样本。可以理解为标签。
- macro average：所有标签结果的平均值。
- weighted average：所有标签结果的加权平均值。

模型优劣的评价指标：

- f1-score：F1分数同时考虑精确率和召回率，让两者同时达到最高，取得平衡。
- precision：精确率，又被称为查准率，是针对预测结果而言的。含义为在被预测为正的样本中实际为正样本的概率。
- recall：召回率，又被称为查全率，是针对原样本而言的。含义为在实际为正的样本中被预测为正样本的概率。
- support：每类标签出现的次数。

图 5-21 测试模型结果



----结束

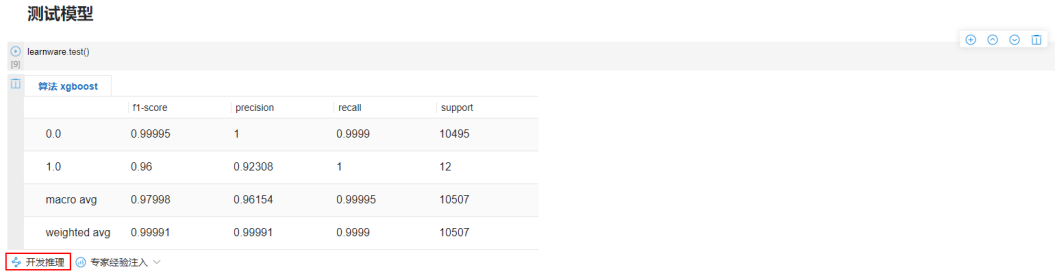
5.4.3.7 开发推理

目前“专家经验注入”是为Gpr数据集定制，如果用户使用Gpr数据集体验KPI异常检测学件的操作流程，可以先执行“专家经验注入”，再执行“开发推理”，那么专家经验会自动转成代码并关联到模型推理函数里面。

“开发推理”用于生成推理代码至推理文件“learnware_predict.py”中。当学件模型打包发布成在线推理服务时，可以使用推理代码，完成快速在线推理验证。推理服务快速验证效果，如**推理服务**所示。

步骤1 单击“测试模型”左下方的“开发推理”，如**图5-22**所示。

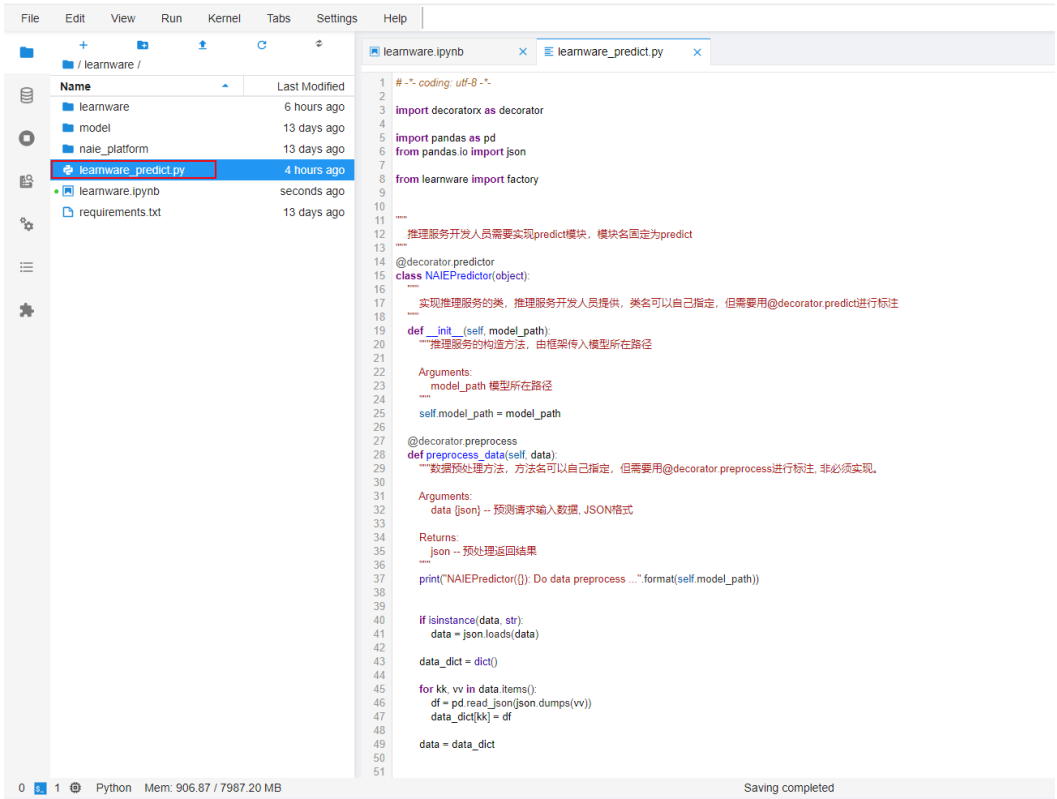
图 5-22 开发推理



步骤2 等待推理代码生成完成后，可在左侧目录树中，看到生成的推理文件“learnware_predict.py”，如**图5-23**所示。

用户可以根据实际情况，编辑修改推理代码。

图 5-23 推理文件



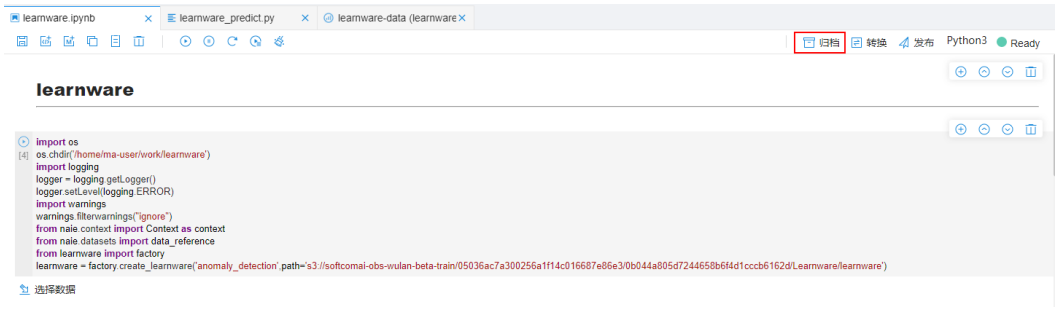
----结束

5.4.3.8 归档模型

模型训练完成后，支持归档模型。操作步骤如下所示。

步骤1 单击界面右上角的“归档”图标，如图5-24所示。

图 5-24 归档图标

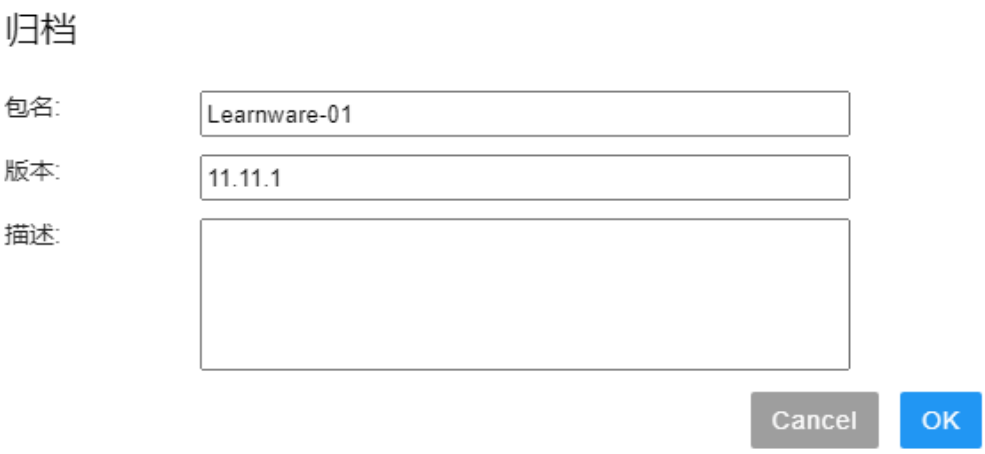


步骤2 在弹出的“归档”对话框中，按照界面提示，配置参数，如图5-25所示。

参数说明如下所示：

- 包名：归档模型的名称。以字母开头，由数字、大小写字母和中划线组成。示例：Learnware-01。
- 版本：归档模型的版本。格式为“xx.yy.zzzz”，其中xx/yy是0-99整数，zzzz为0-9999整数。示例：11.11.1。
- 描述：归档模型的相关描述。请根据实际情况设置。

图 5-25 归档模型



步骤3 单击“OK”，完成模型归档。

----结束

5.4.4 模型管理

在模型管理界面，可以将归档的模型，打包成模型包。

步骤1 在菜单栏中，单击“模型管理”。

进入“模型管理”界面，如图5-26所示。

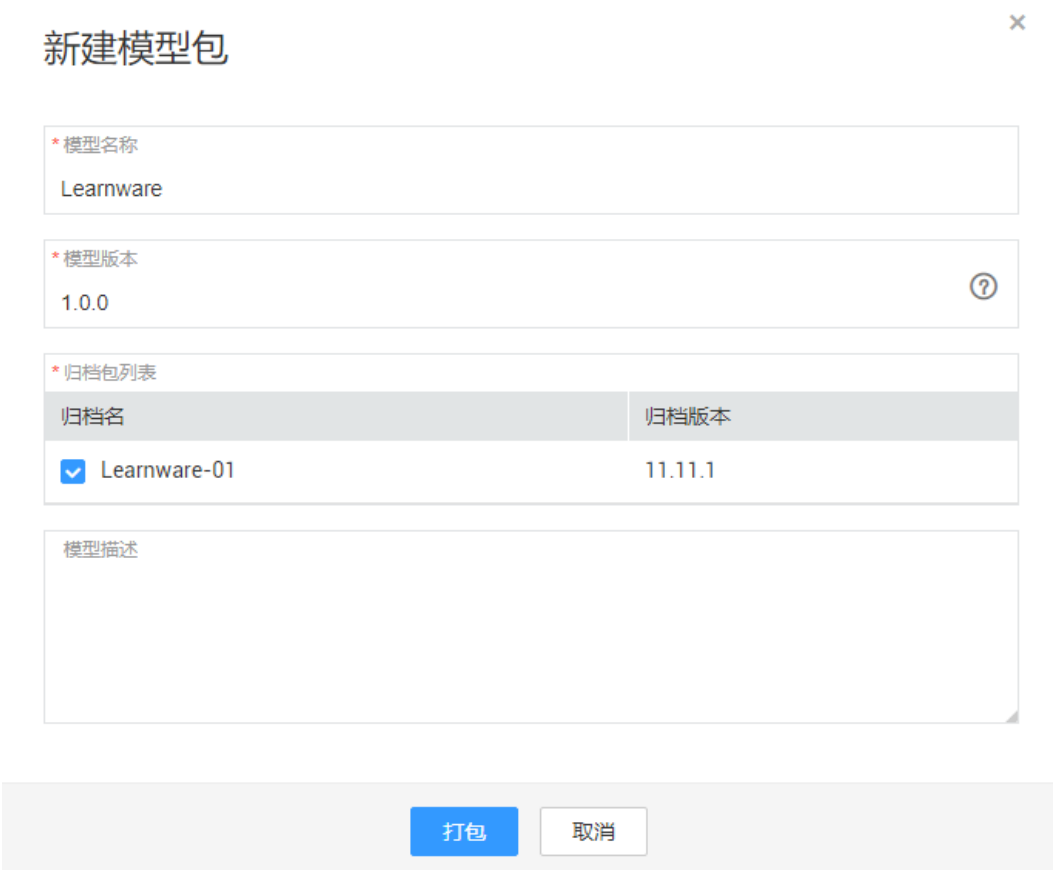
图 5-26 模型管理界面

步骤2 单击界面右上角的“新建模型包”。

弹出“新建模型包”对话框，如图5-27所示。

请根据实际情况，修改模型名称、模型版本、模型描述等信息，并勾选归档的学件模型“Learnware-01”.

图 5-27 新建模型包



步骤3 单击“打包”，将归档的KPI异常检测学件模型打包成模型包。

打包完成后，界面新增“Learnware”模型包，如图5-28所示。

图 5-28 模型包

----结束

5.4.5 推理服务

支持基于模型包，创建推理服务，直接在线调用服务得到推理结果。

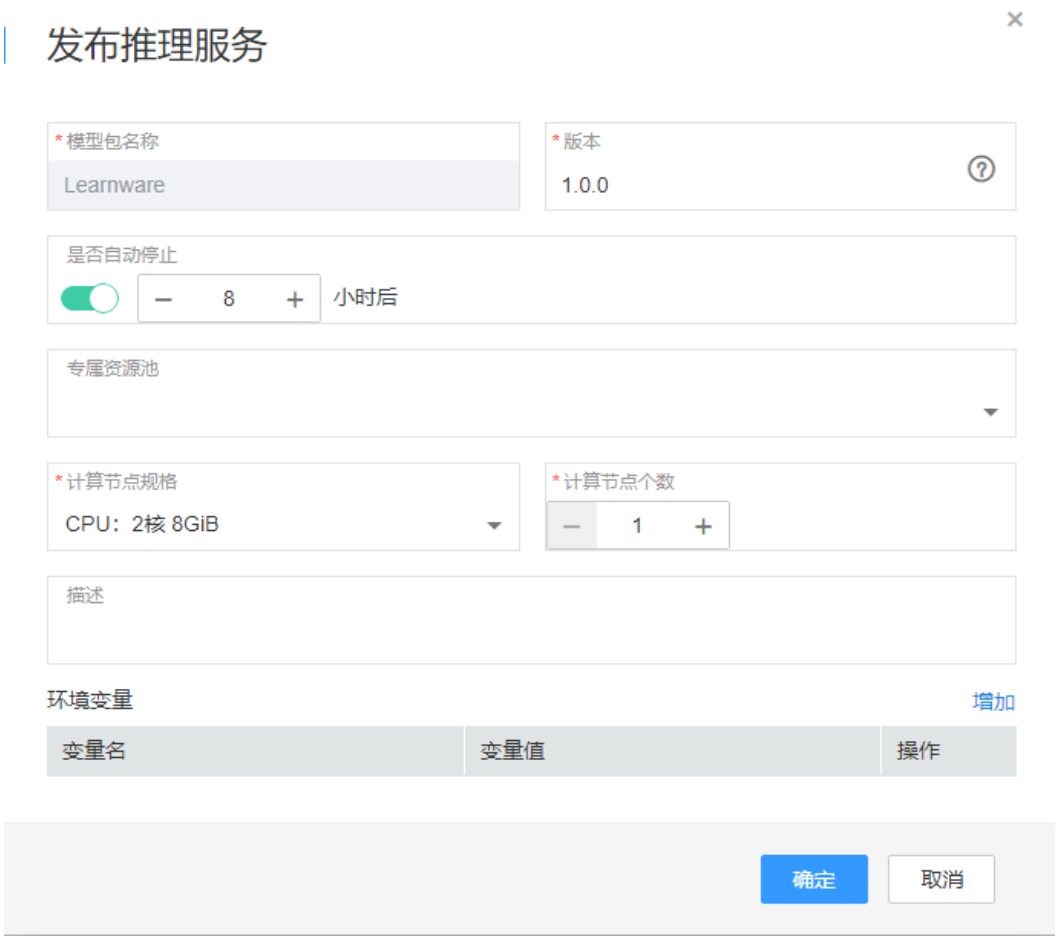
步骤1 在“模型管理”界面，单击学件模型所在行，对应“操作”列的 图标。

弹出“发布推理服务”对话框，如图5-29所示。

请根据实际情况配置如下参数，其余参数保持默认值即可。

- 版本：推理服务的版本。
- 是否自动停止：推理服务的运行时间。建议可以设置长点时间，最长支持24小时。
- 计算节点规格：CPU和GPU资源规格。
- 计算节点个数：“1”代表单节点运算，“2”代表分布式计算。

图 5-29 发布推理服务



步骤2 单击“确定”，发布推理服务。

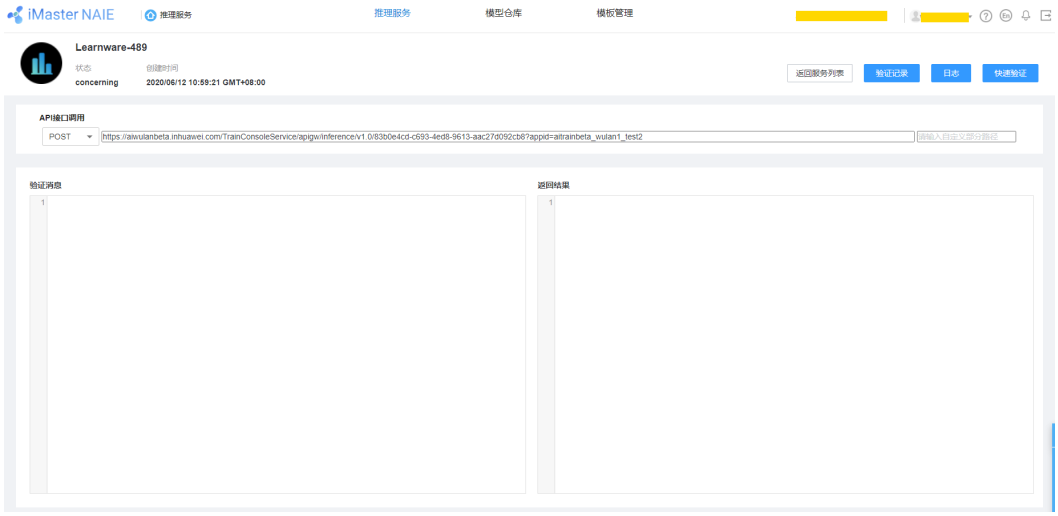
单击“模型管理”界面右上方的“推理服务”链接，可进入云端推理框架的“推理服务”主界面，该界面展示当前发布的所有推理服务。用户可以对推理服务进行查看详情、授权、启动/停止等一系列操作。

-  ：推理服务发布成功，单击图标可以跳转至推理服务的快速验证界面，用户可在此界面上对当前发布的推理服务效果进行验证。
-  ：推理服务发布失败，可重新发布。

步骤3 等待推理服务发布成功后，单击学件模型所在行，对应“操作”列的  图标。

进入推理服务快速验证界面，如图5-30所示。

图 5-30 推理服务快速验证界面



步骤4 在验证信息栏，输入json格式的验证数据。

示例如下所示：

```
{
   "data3.csv": {
      "time": {
         "0": "2018\\/7\\/12 16:28",
         "1": "2018\\/7\\/12 16:29",
         "2": "2018\\/7\\/12 16:30",
         "3": "2018\\/7\\/12 16:31",
         "4": "2018\\/7\\/12 16:32",
         "5": "2018\\/7\\/12 16:33",
         "6": "2018\\/7\\/12 16:34",
         "7": "2018\\/7\\/12 16:35",
         "8": "2018\\/7\\/12 16:36",
         "9": "2018\\/7\\/12 16:37",
         "10": "2018\\/7\\/12 16:38"
      },
      "send_byte": {
         "0": 0,
         "1": 0,
         "2": 0,
         "3": 0,
         "4": 0,
         "5": 0,
         "6": 0,
         "7": 0,
         "8": 0,
         "9": 0,
         "10": 0
      }
   }
}
```

步骤5 单击界面右上角的“快速验证”图标，调用推理服务，返回推理结果。

----结束

5.5 多层嵌套异常检测学件

5.5.1 创建项目

多层嵌套异常检测学件服务，目前封装在模型训练服务的JupyterLab平台中。可通过在项目中创建JupyterLab环境，体验多层嵌套异常检测学件服务。

步骤1 在训练平台首页，单击界面左上角的“创建项目”图标。

弹出“创建项目”对话框，如图5-31所示。

请根据实际情况，配置如下参数：

- 名称：项目名称。
- 是否公开：项目是否可以被所属用户组的其他用户访问。
- 公开至组：仅当“是否公开”设置为“是”，才会展示“公开至组”。默认展示当前用户所属的所有用户组，如果勾选用户所属的用户组，则被勾选用户组下的所有用户均可以查看当前项目。如果只想公开给用户组中的部分用户，可以单击“请选择用户”，筛选出希望共享的用户。

图 5-31 创建项目

步骤2 单击“创建”。项目创建完成后，进入项目概览界面。

----结束

5.5.2 样例数据导入训练平台

步骤1 在项目概览界面，单击菜单栏中的“特征工程”，进入“特征工程”界面。

步骤2 单击界面右上角的“特征处理”，弹出“特征处理”对话框，如图5-32所示。

请根据实际情况，配置如下参数：

- 工程名称：特征工程名称。
- 开发模式：请选择“Jupyterlab交互式开发”。
- 规格：选择Jupyterlab环境部署的容器规格大小。
- 实例：从下拉框中选择“新建一个环境”。

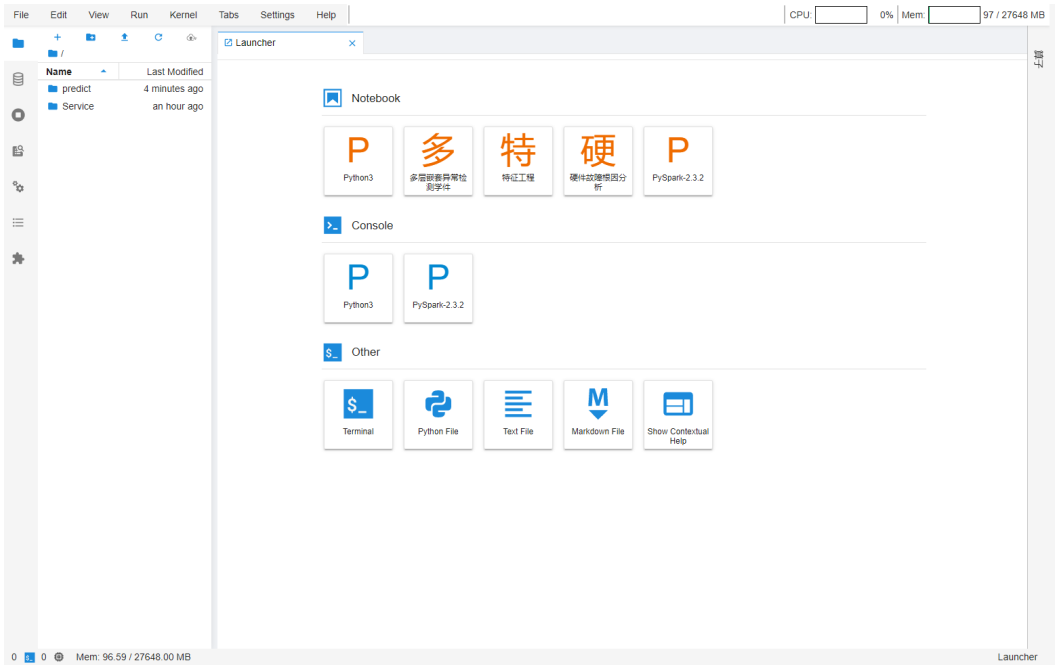
图 5-32 特征处理

步骤3 单击“创建”，等待Jupyterlab环境创建完成，约需要5分钟。

步骤4 等待Jupyterlab环境创建完成后，单击特征工程所行，对应操作列的图标。

进入Jupyterlab环境首页，如图5-33所示。

图 5-33 Jupyterlab 环境首页



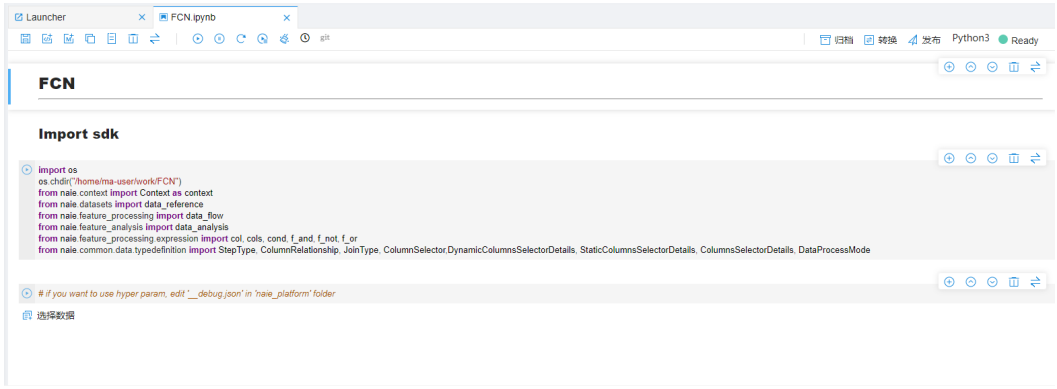
步骤5 单击Notebook区域下方的“多层嵌套异常检测学件”，弹出“新建”对话框。

步骤6 在弹出的“新建”对话框中，输入学件名称，示例为“FCN”，单击“OK”。

进入“FCN.ipynb”文件界面，如图5-34所示。

系统会弹出“Select Kernel”对话框，选择“Python3”，并单击“Select”。

图 5-34 “FCN.ipynb”文件界面



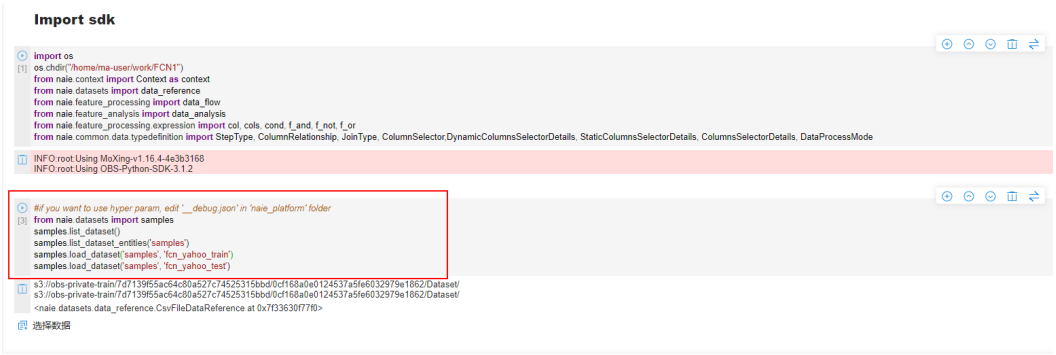
步骤7 单击“Import sdk”代码框左侧的图标，导入算法依赖的训练平台SDK。

步骤8 在如图5-35所示的空白代码框中输入如下代码并运行。

将“samples”数据导入训练平台。

```
#if you want to use hyper param, edit '__debug.json' in 'naie_platform' folder
from naie.datasets import samples
samples.list_dataset()
samples.list_dataset_entities('samples')
samples.load_dataset('samples', 'fcn_yahoo_train')
samples.load_dataset('samples', 'fcn_yahoo_test')
```


图 5-35 导入 samples 数据至训练平台



----结束

5.5.3 模型训练

步骤1 单击代码框左下方的“选择数据”。

弹出“选择数据”代码框，如图5-36所示。

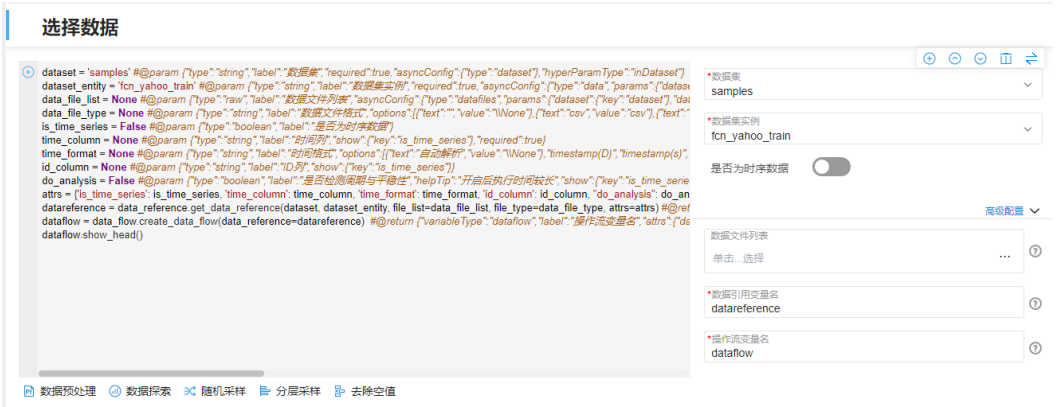
说明

也可以通过界面右上方的菜单“算子 > 学件 > 多层嵌套异常检测学件 > 选择数据”，添加“选择数据”代码框。

参数说明如下所示：

- 数据集：从下拉框中选择数据集“samples”。
- 数据集实例：从下拉框中选择训练数据“fcn_yahoo_train”。

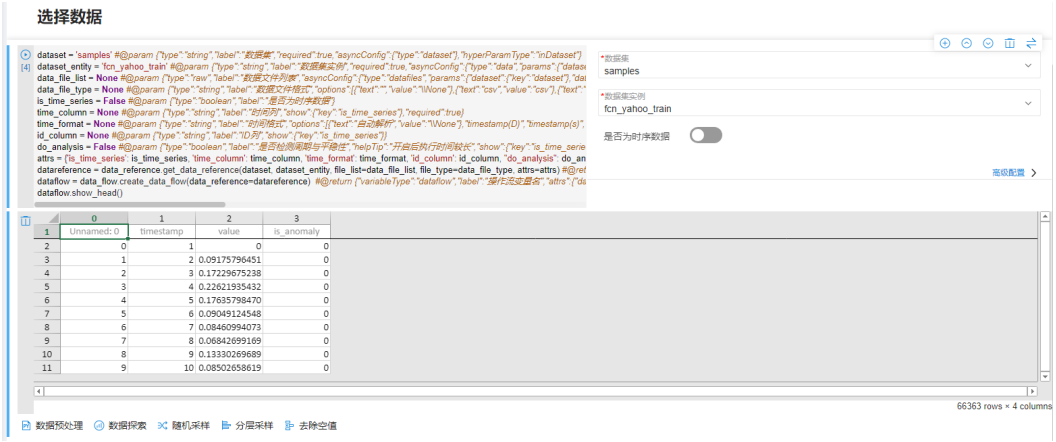
图 5-36 选择数据



步骤2 单击“选择数据”代码框左侧的 图标。运行代码，绑定训练数据。

运行成功后，可以查看训练数据，如图5-37所示。

图 5-37 查看训练数据



步骤3 单击界面左下角的“数据预处理”。

弹出“数据预处理”代码框，如图5-38所示。

说明

也可以通过界面右上方的菜单“算子 > 学件 > 多层嵌套异常检测学件 > 数据预处理”，添加“数据预处理”代码框。

参数说明如下所示：

- 指标列：保持默认值“value”。
- 标签列：保持默认值“is_anomaly”。
- 数据处理模式：保持默认值“训练”。

图 5-38 数据预处理

步骤4 单击“数据预处理”代码框左侧的图标。运行代码，对训练数据做数据预处理。

步骤5 单击界面左下角的“异常检测模型训练”。

弹出“异常检测模型训练”代码框，如图5-39所示。

请根据实际情况配置各个模型参数取值。

说明

也可以通过界面右上方的菜单“算子 > 学件 > 多层嵌套异常检测学件 > 异常检测模型训练”，添加“异常检测模型训练”代码框。

图 5-39 异常检测模型训练

步骤6 单击“异常检测模型训练”代码框左侧的图标。等待模型训练完成。

可以通过屏幕打印信息，查看模型训练过程。屏幕会依次打印400个Epochs的模型训练评估结果。

----结束

5.5.4 模型测试

步骤1 单击界面左下角的“选择数据”。

弹出“选择数据”代码框，如图5-40所示。

说明

也可以通过界面右上方的菜单“算子 > 学件 > 多层嵌套异常检测学件 > 选择数据”，添加“选择数据”代码框。

参数说明如下所示：

- 数据集：从下拉框中选择数据集“samples”。
- 数据集实例：从下拉框中选择训练数据“fcn_yahoo_test”。

展开“高级配置”，配置“数据引用变量名”，因为特征工程引用了多份数据，分别为训练数据和测试数据，为避免冲突，修改测试数据的“数据引用变量名”为“datareference1”。

图 5-40 选择数据

步骤2 单击“选择数据”代码框左侧的图标。运行代码，绑定测试数据。

运行成功后，可以查看测试数据。

步骤3 单击界面左下角的“数据预处理”。

弹出“数据预处理”代码框，如图5-41所示。

说明

也可以通过界面右上方的菜单“算子 > 学件 > 多层嵌套异常检测学件 > 数据预处理”，添加“数据预处理”代码框。

参数说明如下所示：

- 指标列：保持默认值“value”。
- 标签列：保持默认值“is_anomaly”。
- 数据处理模式：选择“测试”。

左侧代码框需要修改如下内容：

- “train_data”在训练数据做数据预处理时已经使用，需要修改为“test_data”。
- “datareference”需要修改为测试数据的引用变量名“datareference1”。

图 5-41 数据预处理

步骤4 单击“数据预处理”代码框左侧的图标。运行代码，对测试数据做数据预处理。

步骤5 单击界面左下角的“异常检测模型测试”。

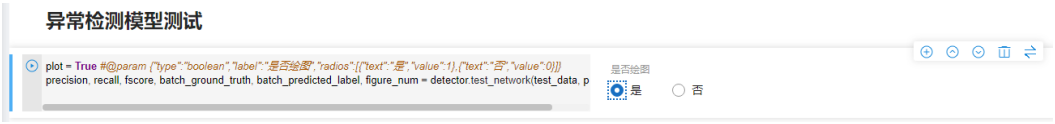
弹出“异常检测模型测试”代码框，如图5-42所示。

“是否绘图”请选择“是”，可以通过绘图查看模型的测试验证效果。

说明

也可以通过界面右上方的菜单“算子 > 学件 > 多层嵌套异常检测学件 > 异常检测模型测试”，添加“异常检测模型测试”代码框。

图 5-42 异常检测模型测试

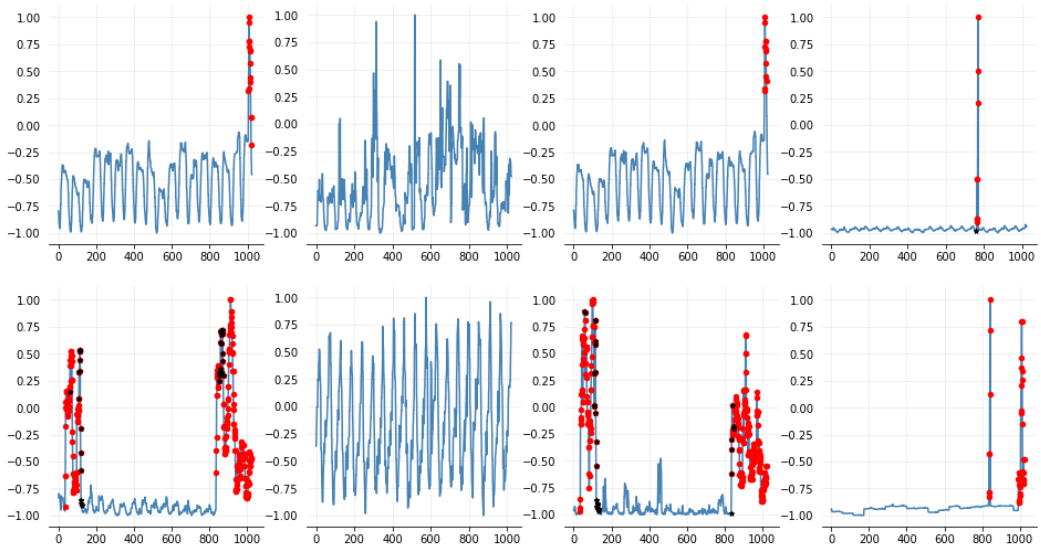


步骤6 单击“异常检测模型测试”代码框左侧的图标。等待模型测试完成。

模型测试打印结果示例，如图5-43所示。截图仅为模型测试打印结果的一部分，具体以实际打印结果为准。

图中黑点是模型预测的异常点，红点是原始异常点。

图 5-43 模型测试结果



----结束

5.6 硬盘故障根因分析学件

5.6.1 创建项目

硬盘故障根因分析学件服务，目前封装在模型训练服务的JupyterLab平台中。可通过在项目中创建JupyterLab环境，体验硬盘故障根因分析学件服务。

步骤1 在训练平台首页，单击界面左上角的“创建项目”图标。

弹出“创建项目”对话框，如图5-44所示。

请根据实际情况，配置如下参数：

- 名称：项目名称。

- 是否公开：项目是否可以被所属用户组的其他用户访问。
- 公开至组：仅当“是否公开”设置为“是”，才会展示“公开至组”。默认展示当前用户所属的所有用户组，如果勾选用户所属的用户组，则被勾选用户组下的所有用户均可以查看当前项目。如果只想公开给用户组中的部分用户，可以单击“请选择用户”，筛选出希望共享的用户。

图 5-44 创建项目

步骤2 单击“创建”。项目创建完成后，进入项目概览界面。

----结束

5.6.2 样例数据导入训练平台

步骤1 在项目概览界面，单击菜单栏中的“特征工程”，进入“特征工程”界面。

步骤2 单击界面右上角的“特征处理”，弹出“特征处理”对话框，如图5-45所示。

请根据实际情况，配置如下参数：

- 工程名称：特征工程名称。

- 开发模式：请选择“Jupyterlab交互式开发”。
- 规格：选择Jupyterlab环境部署的容器规格大小。
- 实例：从下拉框中选择“新建一个环境”。

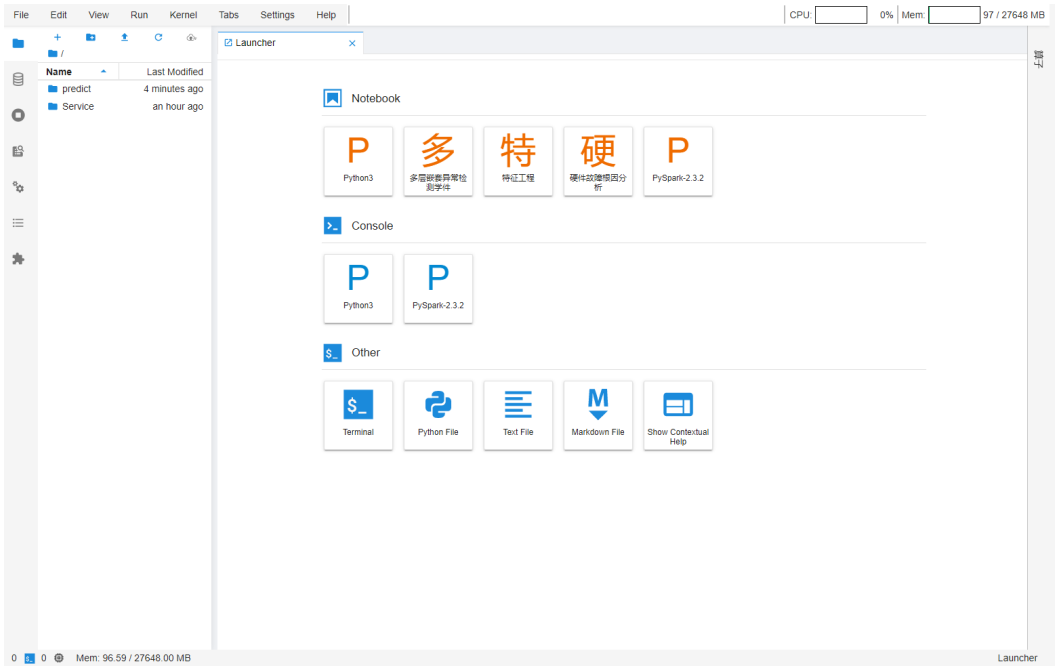
图 5-45 特征处理

步骤3 单击“创建”，等待Jupyterlab环境创建完成，约需要5分钟。

步骤4 等待Jupyterlab环境创建完成后，单击特征工程所行，对应操作列的图标。

进入Jupyterlab环境首页，如图5-46所示。

图 5-46 Jupyterlab 环境首页



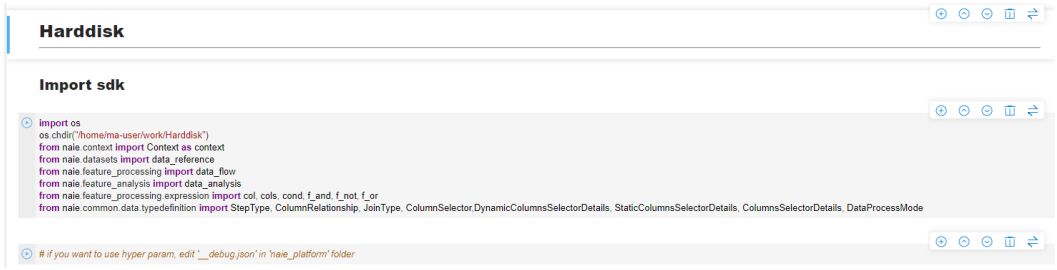
步骤5 单击Notebook区域下方的“硬盘故障根因分析学件”，弹出“新建”对话框。

步骤6 在弹出的“新建”对话框中，输入学件名称，示例为“Harddisk”，单击“OK”。

进入“Harddisk.ipynb”文件界面，如图5-47所示。

系统会弹出“Select Kernel”对话框，选择“Python3”，并单击“Select”。

图 5-47 “Harddisk.ipynb”文件界面



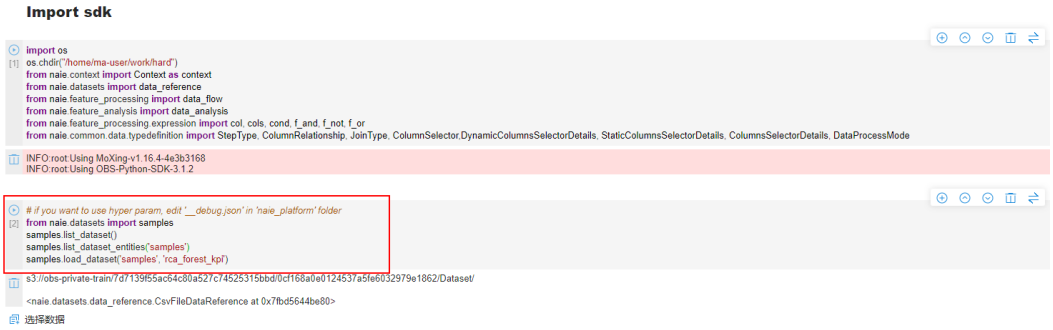
步骤7 单击“Import sdk”代码框左侧的图标，导入算法依赖的训练平台SDK。

步骤8 在如图5-48所示的空白代码框中输入如下代码并运行。

将“samples”数据导入训练平台。

```
# if you want to use hyper param, edit '__debug.json' in 'naie_platform' folder
from naie.datasets import samples
samples.list_dataset()
samples.list_dataset_entities('samples')
samples.load_dataset('samples', 'rca_forest_kpi')
```


图 5-48 导入 samples 数据至训练平台



----结束

5.6.3 模型训练

步骤1 单击代码框左下方的“选择数据”。

弹出“选择数据”代码框，如图5-49所示。

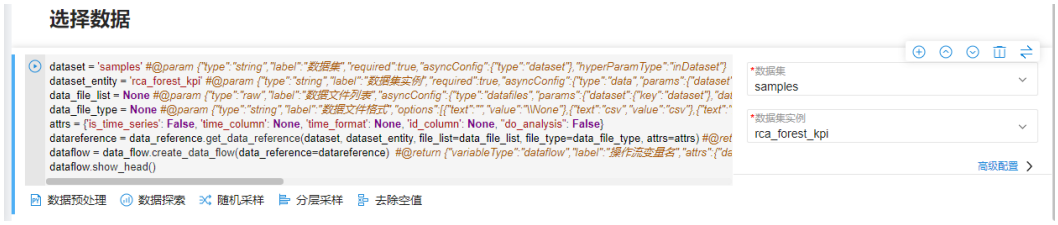
说明

也可以通过界面右上方的菜单“算子 > 学件 > 硬盘故障根因分析 > 选择数据”，添加“选择数据”代码框。

参数说明如下所示：

- 数据集：从下拉框中选择数据集“samples”。
- 数据集实例：从下拉框中选择数据“rca_forest_kpi”。

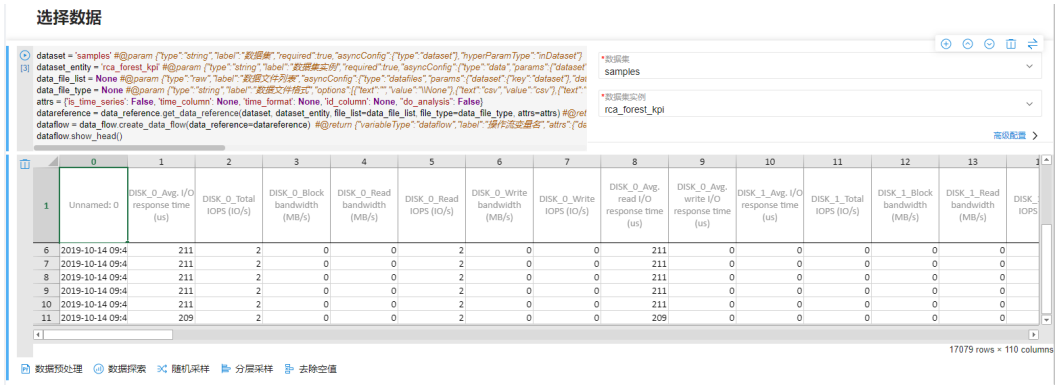
图 5-49 选择数据



步骤2 单击“选择数据”代码框左侧的 图标。运行代码，绑定数据。

运行成功后，可以查看数据，如图5-50所示。

图 5-50 查看训练数据



步骤3 单击界面左下角的“数据预处理”。

弹出“数据预处理”代码框，如图5-51所示。

说明

也可以通过界面右上方的菜单“算子 > 学件 > 硬盘故障根因分析 > 数据预处理”，添加“数据预处理”代码框。

参数说明如下所示：

- 列筛选方式：保持默认值“列选择”。
- 待处理列：选择除时间列“Unnamed: 0”外的所有列。
- 时间列：选择时间列“Unnamed: 0”。
- 分组数：请根据实际业务场景配置分组数量。如果配置为“2”，数据预处理后的效果如图5-52所示，将相邻两行数据合并为一行展示。如果相邻四行的数据具备相关性，则需要将4行数据合并为一行展示，“分组数”配置为“4”。此处，保持默认值“2”。
- 标签列：选择标签列“label”。
- 标签汇聚方式：取值如果为“logic_or”，则转换后的标签列值为转换前的多个标签列值做逻辑或运算；取值如果为“logic_and”，则转换后的标签列值为转换前的多个标签列值做逻辑与运算。

图 5-51 数据预处理

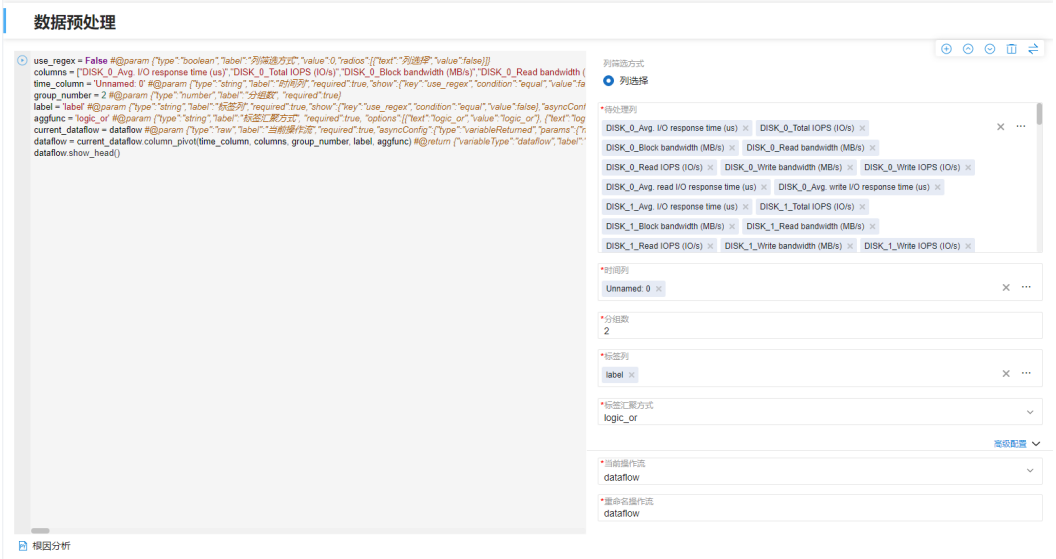


图 5-52 分组后的数据转换效果

time	id	value
2020/6/15 08:00:00	1	0.5
2020/6/15 08:05:00	1	0.6
2020/6/15 08:10:00	1	0.7
2020/6/15 08:15:00	1	0.8

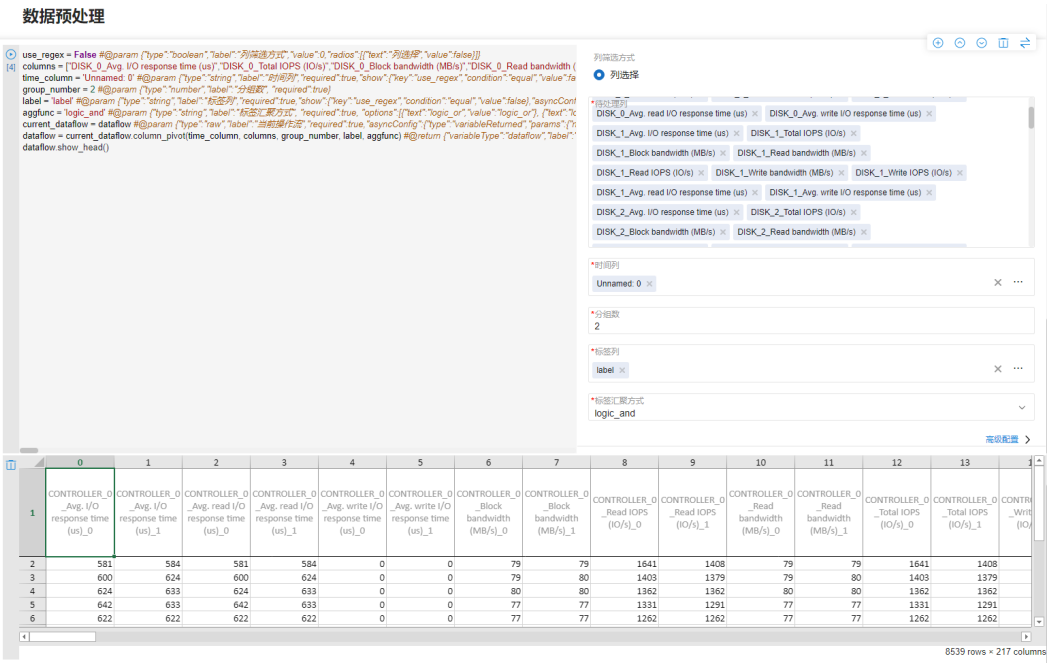
column_pivot

time	value_0	value_1
2020/6/15 08:05:00	0.5	0.6
2020/6/15 08:15:00	0.7	0.8

步骤4 单击“数据预处理”代码框左侧的图标。运行代码，对数据做数据预处理。

数据预处理后的结果，如图5-53所示。

图 5-53 数据预处理结果



步骤5 单击界面左下角的“根因分析”。

弹出“根因分析”代码框，如图5-54所示。

请根据实际情况配置模型参数取值。当前支持使用RandomForest、XGBoost、使用RandomForest和XGBoost的Ensemble三种算法的模型进行特征评估。其中设置的“选择根因数”的值，为在“结果展示”的运行结果图中展示的根因KPI个数。

说明

也可以通过界面右上方的菜单“算子 > 学件 > 硬盘故障根因分析 > 根因分析”，添加“根因分析”代码框。

图 5-54 根因分析

步骤6 单击“根因分析”代码框左侧的 图标。等待根因分析完成。

步骤7 单击界面左下角的“结果展示”。

弹出“结果展示”代码框。

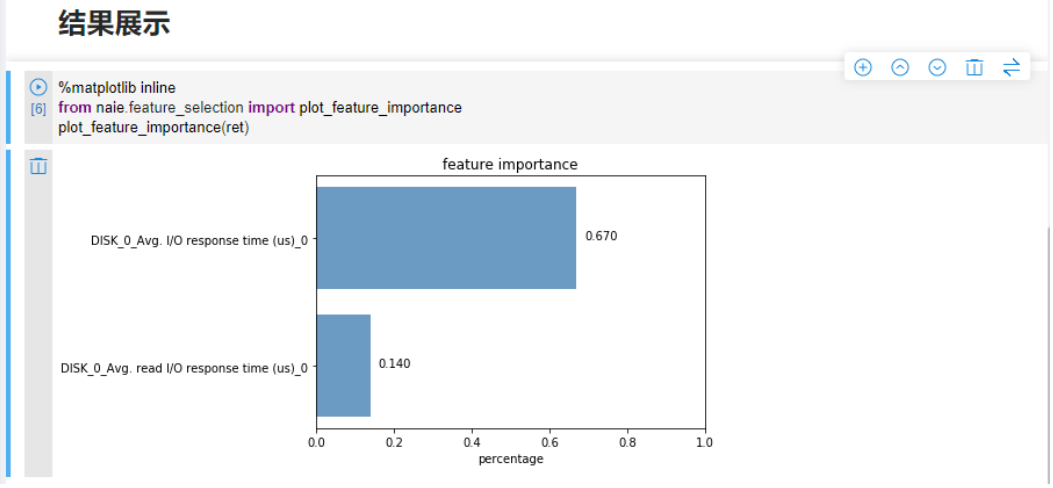
说明

也可以通过界面右上方的菜单“算子 > 学件 > 硬盘故障根因分析 > 结果展示”，添加“结果展示”代码框。

步骤8 单击“结果展示”代码框左侧的图标。

运行完成后，效果如图5-55所示。可以通过结果图，查看模型推荐的造成硬盘故障的前两个根因KPI和占比情况。

图 5-55 结果展示



----结束

5.7 修订记录

发布日期	修订记录
2019-07-30	新增多层嵌套异常检测学件，对应新增“**多层嵌套异常检测学件**”。 新增硬盘故障根因分析学件，对应新增“**硬盘故障根因分析学件**”。
2019-06-30	第一次正式发布。

6 常见问题

6.1 训练平台首页

6.1.1 如何回到训练平台首页?

用户进入项目总览界面、数据集页面、特征工程页面、模型训练页面、模型管理页面或者模型验证界面后，如果需要回到项目列表首页，请单击界面左上角品牌名称右侧的服务名称，从下拉框中选择“模型训练服务”。

6.1.2 创建项目公开至组的参数是什么含义?

用户在创建IAM用户的时候会涉及到用户组的概念。将IAM用户加入指定的用户组，则IAM用户和此用户组的账号权限相同。

创建项目时选择的公开至组，这个组即是当前IAM用户所属的用户组。勾选用户组后，此组内所有的IAM用户都可以看到当前IAM用户所创建的项目，进行经验共享、协同工作。

6.2 特征工程

6.2.1 如何选中全量特征列?

使用Python和Spark开发平台创建的特征工程，在特征操作界面，单击表格左上方第一个带有倒三角标识的单元格即可。

使用JupyterLab开发平台创建的特征工程，在特征工程操作编辑区域分别运行“Import sdk”和“绑定数据”代码框。运行成功后，在“绑定数据”代码框下方单击全量特征表格左上方第一个带有倒三角标识的单元格即可。

6.2.2 特征工程处理的时候必须要先采样吗?

特征工程数据采样的目的是提升界面每个特征操作的速度。大数据量操作的时候建议先采样。数据采样后所有的特征操作，都只对采样后的数据进行处理，可以减少特征操作处理的数据量。

6.2.3 特征处理操作完成后怎么应用于数据集全量数据?

使用Python和Spark开发平台创建的特征工程，界面所有特征操作执行完成后，单击“执行”时，系统自动将特征操作流应用于数据集全量数据，生成经过特征处理的数据集，供模型训练使用。用户在单击“执行”时，可以在“执行”对话框中，选择其他数据集，执行当前的特征操作流。添加的数据集，必须满足特征维度和特征列数量与当前特征工程绑定的数据集一致，否则会执行失败。

使用JupyterLab开发平台创建的特征工程，界面所有特征操作执行完成后，在界面右上角选择“算子 > 数据处理 > 数据集 > 生成数据实例”，在新增的“生成数据实例”代码框右侧选择数据集和数据实例，运行代码框。系统自动将特征操作流应用于数据集全量数据，生成经过特征处理的数据集，供模型训练使用。

用户可以在“数据集”界面查看新生成的数据。

6.3 模型训练

6.3.1 模型训练新建模型训练工程的时候，选择通用算法有什么作用?

通用算法目前包括：分类算法、拟合算法、聚类算法、其他类型。用户选择不同的通用算法类型，并勾选“创建入门模型训练代码”，便可以自动生成对应类型的代码模版。

6.3.2 使用训练模型进行在线推理的推理入口函数在哪里编辑?

进入简易编辑器界面，在“代码目录”节点下，创建推理文件，根据实际情况写作推理代码。

6.3.3 通过数据集导入数据后，在开发代码中如何获取这些数据?

训练平台提供了SDK供开发人员直接获取数据集，具体使用方式如下所示：

步骤1 导入训练平台SDK。

```
from naie.datasets import data_reference
from naie.feature_processing import data_flow
```

步骤2 使用get_data_reference获取数据集存放路径。

以数据集“air”、数据集实例“air_20190409”为例，此时SDK返回的是数据集所存储文件路径。

```
data_reference=get_data_reference(dataset="air",dataset_entity="air_20190409")
```

----结束

6.3.4 如何在模型训练时，查看镜像中 Python 库的版本?

在训练的代码中增加如下所示的代码行，执行训练即可查看：

```
print(os.system("pip list"))
```

6.3.5 如何在模型训练时，设置日志级别?

在TensorFlow的log日志等级如下：

- - 0：显示所有日志（默认等级）
- - 1：显示info、warning和error日志
- - 2：显示warning和error信息
- - 3：显示error日志信息

以设置日志级别为“3”为例，操作方法如下：

```
os.environ['TF_CPP_MIN_LOG_LEVEL']='3'
```

6.3.6 如何自定义安装 python 第三方库?

如何在训练平台中安装算法依赖的库，方法如下所示：

- 训练服务支持使用pip安装，算法依赖的第三方库，以安装pystan为例，操作方法如下：
  ```
  os.system("pip install pystan")
  ```
- Notebook支持使用pip安装，算法依赖的第三方库，以安装pystan为例，操作方法如下：
  ```
  !pip install pystan == 1.0.0
  ```
- 训练服务和Notebook支持使用requirements.txt安装，算法依赖的第三方库。“requirements.txt”文件仅支持安装pip仓库中已有的包，否则安装失败。以安装pystan为例，操作方法如下：
  ```
  pystan == 1.0.0
  ```

6.4 模型验证

6.4.1 模型验证服务是什么含义?

模型验证界面支持创建验证服务，并编辑模型验证算法。加入验证时，需要选择已经打包好的模型，设置AI引擎、验证数据集、验证数据集实例、标签列、运行参数、计算节点规格。验证完成后，查看验证报告中模型的准确率等信息。

6.5 通用问题

6.5.1 AutoML 的使用入口有哪些?

包含如下入口：

1. 在“特征工程”菜单界面，创建JupyterLab环境的方式。在JupyterLab界面，选择右上角的“算子 > 模型训练 > 模型训练 > AutoML”，新增AutoML内容，实现零编码使用AutoML。
2. 在“模型训练”菜单界面，创建WebIDE环境的方式。在WebIDE中导入AutoML模块，代码为“from naie.automl import VegaAutoML”。通过代码调用SDK的方式，便于与其他代码的集成开发和调试。
3. 通过提交模型训练任务的方式。因为AutoML一般需要很多次迭代过程，且运行时间很长。为了运行多个任务，模型训练服务提供提交训练任务的方式，运行AutoML。

6.6 修订记录

发布日期	修订记录
2020-08-30	新增“**AutoML的使用入口有哪些？**”章节。 更新如下章节内容： • **特征工程处理的时候必须要先采样吗？** • **特征处理操作完成后怎么应用于数据集全量数据？** • **使用训练模型进行在线推理的推理入口函数在哪里编辑？** • **通过数据集导入数据后，在开发代码中如何获取这些数据？**
2020-03-30	本次版本无变更。
2019-12-30	根据训练平台的菜单，对问题进行分类。
2019-10-30	新增如下章节： • **如何回到训练平台首页？** • **通过数据集导入数据后，在开发代码中如何获取这些数据？** • **如何在模型训练时，查看镜像中Python库的版本？** • **如何在模型训练时，设置日志级别？** • **如何自定义安装python第三方库？**
2019-04-30	第一次正式发布。

7 产品术语

A

AI应用市场

提供AI模型的交易市场，是AI消费者接触NAIE云服务的线上门户，是AI消费者对已上架的AI模型进行查看、试用、订购、下载和反馈意见的场所。

AI引擎

可支持用户进行机器学习、深度学习、模型训练作业开发的框架，如Tensorflow、Spark MLlib、MXNet、PyTorch等。

B

标签列

模型训练输出的预测值，对应数据集的一个特征列。例如鸢尾花分类建模数据集提供了五列数据：花瓣的长度和宽度、花萼的长度和宽度、鸢尾花种类。其中鸢尾花种类就是标签列。

C

超参

模型外部的参数，必须用户手动配置和调整，可用于帮助估算模型参数值。

M

模型包

模型训练完成后，归档或打包模型展示在“模型管理”界面。可以基于模型包创建模型验证服务、训练服务。可以上架至应用市场，支持用户订购后，下载并部署至推理框架中使用。可以一键发布成在线推理服务，一键创建联邦学习实例。也可以对下载的模型包进行完整性校验。

N

Notebook

交互式记事本。用于编写代码的环境。用户可使用R、Python、Scala和SQL等语言编写代码。

P

Python语言

一种可移植、解释性、面向对象的程序设计语言，开发者开发出来并将其免费分发。Python可以运行在许多平台上，包括UNIX、Windows、OS/2、Macintosh等操作系统，可以用来编写TCP/IP应用程序。

S

数据采样

在其他特征操作前先对数据集进行样本采样。数据采样后所有的特征操作，都是基于采样后的数据进行处理，可以减少特征操作处理的数据量，提升特征操作的处理速度。

数据服务

支持网络工参、性能、告警等各种类型数据的快速采集。一方面提供大量工具提升数据治理效率，同时应用多租户隔离、加密存储等安全技术，保障数据的全生命周期安全。

数据集

某业务下具有相同数据格式的数据逻辑集合。

数据集实例

数据集的实例，有具体的数据。

数据准备

数据集中导入的数据实例，可能存在空值、冗余、数据不足等情况，或者用户需要进行数据连接、数据联合、数据修复等操作。

在旧版体验式开发模式下，数据准备包含的功能有：数据修复、数据过滤、数据联合、数据连接、数据去噪。对应JupyterLab交互式开发模式界面“算子 > 数据处理”菜单下面的部分数据处理项。

Schema

用于描述点或边的属性信息，Schema由多个标签组成，每个标签又由1个或多个属性组成。

T

特征操作

特征操作主要是对特征的样本数据值进行修改，也可以重命名、删除、筛选特征列。

在旧版体验式开发模式下，训练平台支持的特征操作有重命名、归一化、数值化、标准化、特征离散化、One-hot编码、数据变换、删除列、选择特征、卡方检验、信息熵、新增特征、PCA。对应JupyterLab交互式开发模式界面“算子 > 数据处理”菜单下面的部分数据处理项。

W

网络AI框架

网络AI框架根据业务场景，可部署在嵌入式网元、网管系统或云侧（私有云或公有云），与不同层级网络控制系统对接，实时采集业务数据，基于最优算法模型实时调整网络运行配置，针对故障实施自动隔离与自动修复，大幅提升网络使用效率与维护效率。

X

训练平台

训练服务为开发者提供电信领域一站式模型开发服务，从数据预处理，到特征提取、模型训练、模型管理、模型验证，本服务为开发者提供开发环境、模拟验证环境，API和一系列开发工具，帮助开发者快速高效开发电信领域模型。

训练数据集

用于训练模型的数据集实例。

Y

验证数据集

模型验证的数据集。

