数据分析挖掘

产品文档

文档目录

- 产品简介
  - 数据挖掘产品概述
    - 词汇表
      - 词汇表
    - 相关服务
    - 相关特性
    - 应用场景
    - 产品优势
- 快速入门
- 操作指南
  - 实验管理
  - 数据集管理
  - 数据集添加
  - 数据集分享
  - 工作流管理
  - 模型管理
  - 推理服务
  - 审批管理
  - 任务中心
  - 全局配置
  - 最佳实践
  - 故障指南
- 常见问题
  - 产品介绍常见问题
  - 产品性能常见问题
  - 产品使用常见问题

产品简介
数据挖掘产品概述
词汇表
词汇表

最近更新时间: 2019-10-28 07:25:26

名词	解释
实验	为了一个目的而进行数据挖掘任务称为实验，一个实验可以使用多种不同的算法或者同一个算法的不同参数组合来实现。
数据集	数据集是对于数据表的抽象，对应于一个对象存储文件或者数据库的一张数据表，数据挖掘组件提供数据读写SDK对数据集进行读写操作。
SDK	数据挖掘组件提供的软件开发工具，可以用代码来实现数据挖掘多种功能。
模型开发环境	开发模型的编程环境，数据挖掘模块提供Jupyternotebook作为模型开发环境。
模型	数据挖掘算法从训练数据中学习到目标函数。在数据挖掘组件中，模型包含模型序列化文件或模型权重文件、模型类文件、模型报告文件、模型超参数、模型指标及模型校验文件等文件。
推理服务	使用数据挖掘模型对数据样本进行预测的服务，包含实时推理服务和离线推理服务。
镜像	Docker镜像是一个文件，由多个层构成，用于在Docker容器内运行代码。镜像本质上是依赖于宿主机OSkernel，用于构建完整可执行应用的命令。数据挖掘组件中，利用Docker镜像来构建不同的模型开发环境。
资源	资源是指与模型构建相关的资源，包含数据集、Notebook、工作流和推理服务等。
模型类文件	模型类文件即Model.py，用于封装模型文件夹，将模型文件夹发布为实时推理和离线推理。
模型指标	用于评估模型效果的指标，根据模型的任务不同可以用不同的指标对模型进行评估，对于分类任务通常用准确率、召回率、AUC等指标；对于回归任务则会选用R方，RSME作为评估指标。

相关服务

最近更新时间: 2019-11-26 15:30:16

- 您可以使用使用数据服务API网关发布具有限流限频功能的推理服务。
- 您可以使用数据管理管理您的项目数据集权限和公开数据集权限。
- 您可以使用智能调度定时调度模型重训练任务。

相关特性

最近更新时间: 2019-11-26 15:30:16

- 数据集管理：异构数据管理，可以将外部的异构数据源添加为抽象的数据集，并且提供统一的程序接口SDK对异构数据进行读写，用于数据挖掘任务和离线推理任务。

- 模型开发：提供可视化建模和Jupyter notebook两种方式进行模型开发，支持主流数据挖掘框架，sklearn、lightGBM、XGBoost、Spark MLlib，以及主流的数据挖掘算法，SVM、逻辑回归、线性回归、决策树、随机森林、协同过滤满足多种数据挖掘需求。

- 模型管理：通过平台训练的模型可以自动进行模型管理，用户可以对相同实验下的模型进行对比，并把任意模型发布为实时推理服务和离线推理服务。模型信息包括：模型序列化文件、模型封装程序（Model.py）、模型训练脚本（Train.py）、模型的超参数信息、模型的指标以及模型校验信息等。

- 推理服务：平台的管理的模型可以通过推理服务，发布为对外的推理服务，包含实时在线推理和离线推理服务。实时推理服务支持自动扩缩容，通过水平扩展支持高并发实时访问。离线推理服务支持TB级数据的离线推理，在集群场景下，Spark框架会进行数据的水平的拆分和合并，在单机场景下，平台会实现数据拆分和合并，用户无需在封装脚本中进行处理。

应用场景

最近更新时间: 2019-11-26 15:30:16

- 金融风控 风险控制：是指风险管理者采取各种措施和方法,消灭或减少风险事件发生的各种可能性,或者减少风险事件发生时造成的损失。金融风控是风险控制在金融领域中的应用。金融领域中需要是使用各种数据挖掘技术和手段对金融事件（信用卡消费行为或贷款申请）或金融主体（个体或企业机构）出现违约的风险进行预测，在金融行业有广泛的应用。

- 营销响应建模 响应建模是预测性数据挖掘技术的主要领域之一，易于实施部署，以获得营销ROI的提升。 响应建模通过定位那些更可能对特定优惠、营销活动、广告、媒体或优惠反应的消费者来改善消费者响应率。这意味着需要通过数据挖掘的手段和技术对每个客户的潜在响应概率进行估计。营销人员将营销预算集中于那些可能响应的受众，而不是全部受众。响应建模是营销人员用较少预算获得更好营销效果的致胜法宝。

- 推荐引擎 推荐引擎，是主动发现用户当前或潜在需求的定律，并主动推送信息给用户的信息网络。挖掘用户的喜好和需求，主动向用户推荐其感兴趣或者需要的对象。 推荐引擎不是被动查找，而是主动推送；不是独立媒体，而是媒体网络；不是检索机制，而是主动学习。推荐引擎利用基于内容、基于用户行为、基于社交关系网络等多种方法，为用户推荐其喜欢的商品或内容。 推荐引擎已经成为大量电商、社交媒体、内容媒体提高用户体验和用户粘性的重要工具。数据挖掘组件内置了多种推荐算法，可以帮助客户搭建自有推荐引擎。

- 流失预警 流失预警是CRM（客户关系管理）中主要技术之一，现代CRM理论会将客户与品牌的关系处在一个生命周期中，不同的生命周期阶段需要采取不同的手段来强化客户与品牌的关系。老客户与品牌的联系变弱的末期通常需要一些挽回手段来对老客户进行挽留和激活。这时候就需要流式预警技术来预测哪些老客户处在流式边缘，以便品牌对这部分客户进行挽留和激活。

- 时间序列分析 对于一些与时间关联紧密的变量或外部关联变量过于复杂的情况，可以根据历史时间的变化情况对变量的未来值进行预测，这种技术被称为时间序列分析，例如，股票价格分析、经济指数分析等。

产品优势

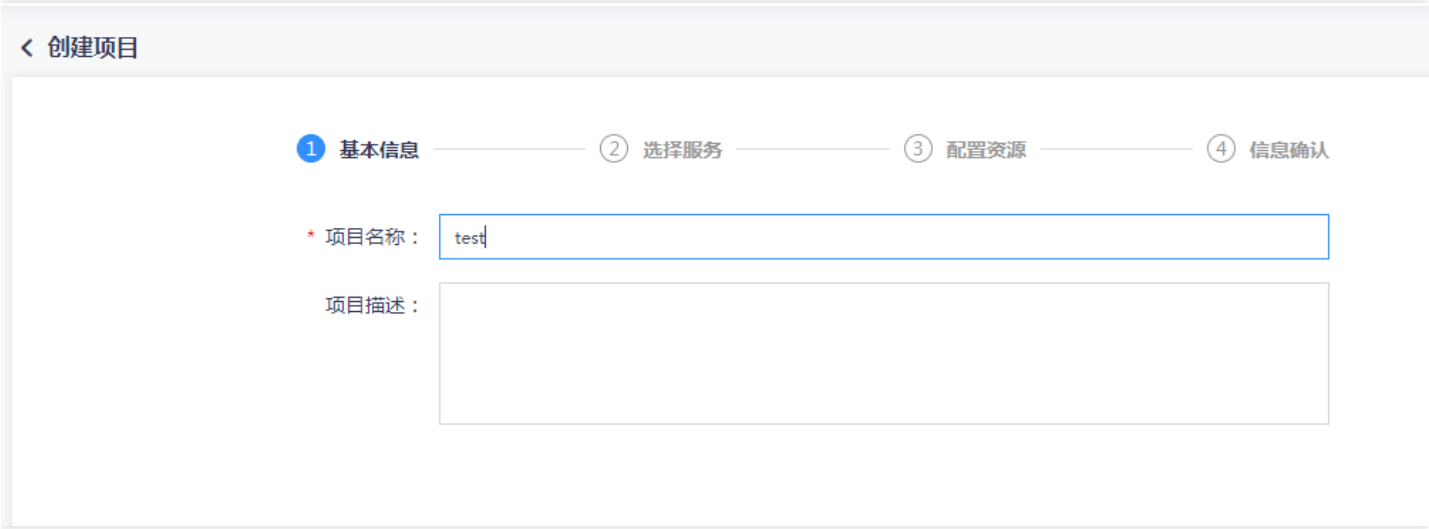
最近更新时间: 2019-11-26 15:30:16

- 丰富算法框架支持 支持主流数据挖掘算法，涵盖分类、聚类、回归等；支持多种数据挖掘框架，sklearn、lightGBM、XGBoost、SparkMLlib等。
- 多种交互方式 提供可视化的方式构建模型训练工作流以及基于notebook来训练模型的方式。 提供了网格搜索（grid-search）和交叉验证（cross-validation）等超参优化策略进行参数调优。
- 一站式数据挖掘平台 覆盖数据挖掘全流程，包含数据读取、特征工程、超参数调优、模型训练、模型部署以及模型重训练等数据挖掘全生命周期。
- 多种推理方式支持 提供了模型封装程序标准，基于标准封装的模型可以一键发布为支持高吞吐、低延时的实时在线推理（基于微服务架构），也可以发布为支持TB级别数据的离线推理。

快速入门

最近更新时间: 2019-11-13 09:20:36

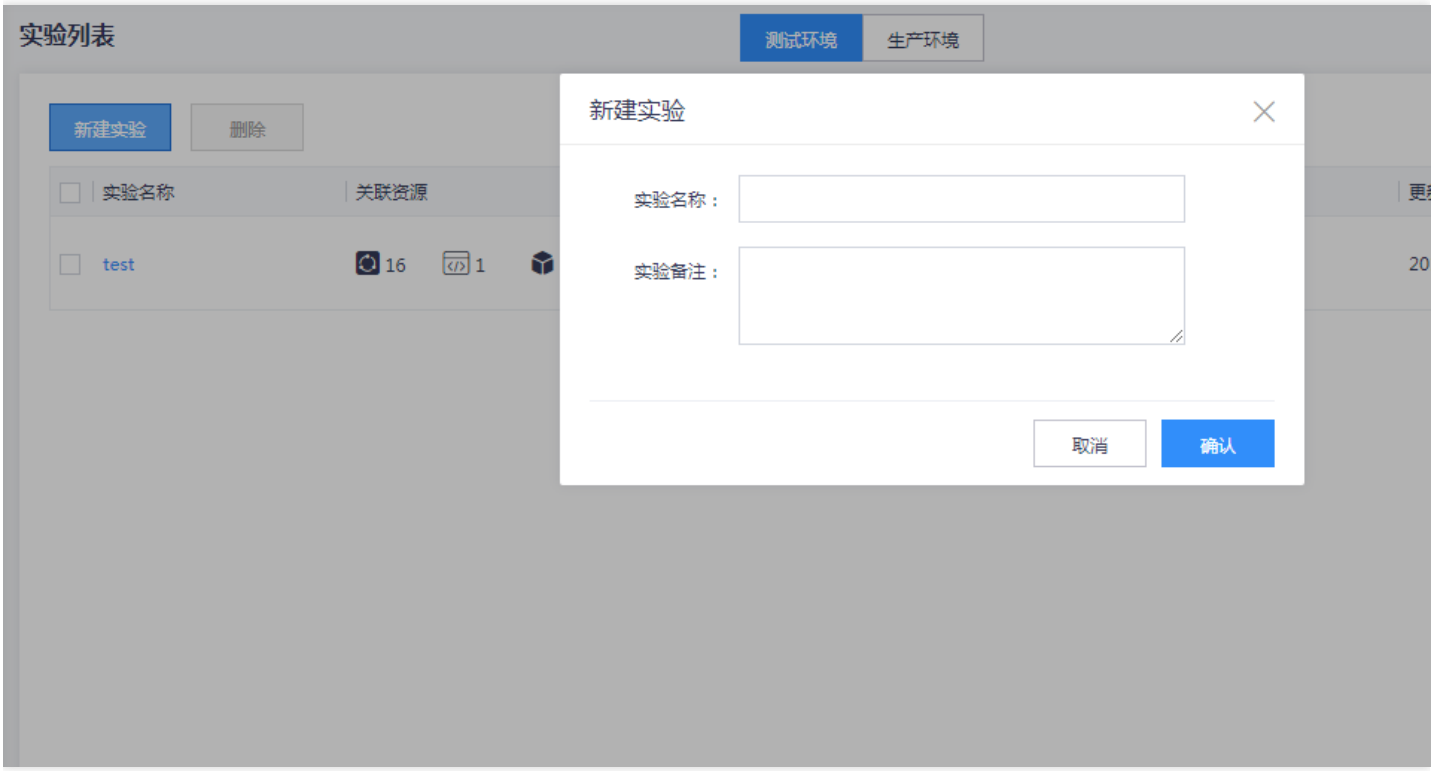
下面将带您体验从开通数据挖掘服务到模型开发、发布推理服务的过程： 1、开通数据分析挖掘服务 2、创建一个项目，填写项目名称



选择数据挖掘服务

选择项目使用的资源组

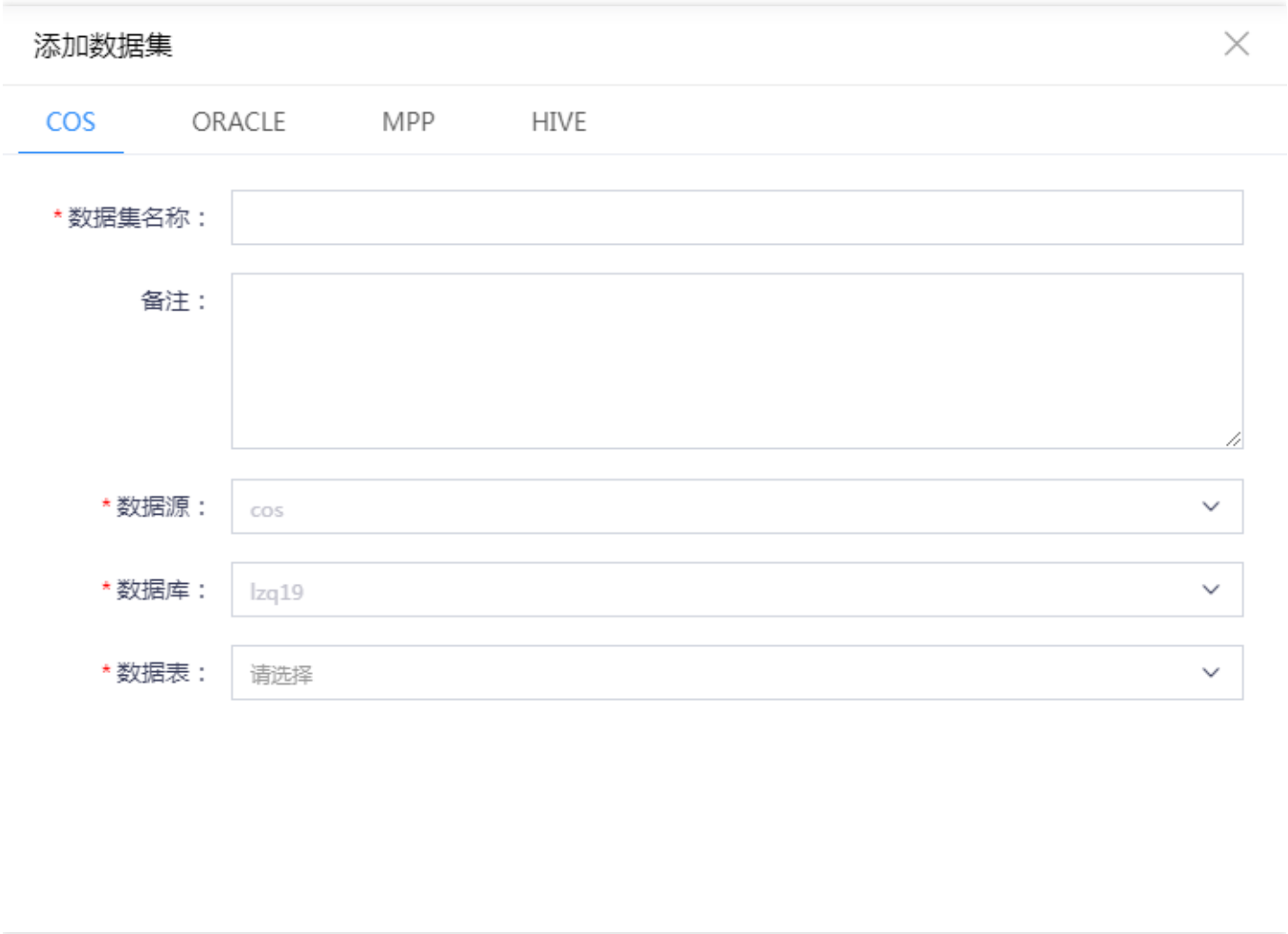
3、新建一个实验，在实验列表中找到新建的实验



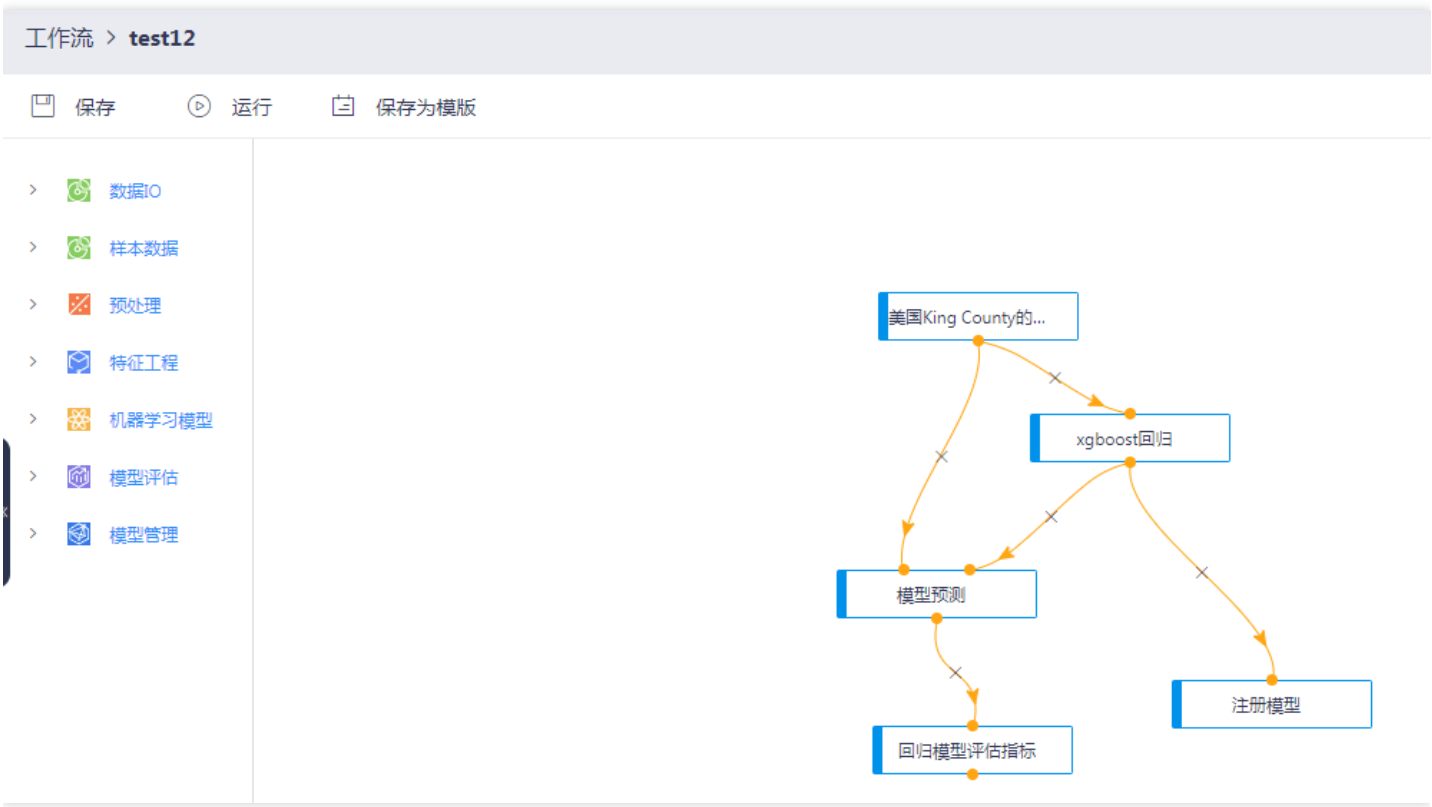
4、新建数据集 数据挖掘组件中有四类数据集：我的数据、收藏的数据、项目数据和公开数据。

我的数据有两种数据集添加方式： a).从本地上传文件到cos b).从数据管理已有的数据表添加

从本地上传数据，需要指定数据集名称、数据文件编码方式及列分隔符。同时需要指定上传的数据源，以及数据的元数据信息。上传文件后还可以添加对应的数据标签。



从数据管理添加，可以将数据管理已有的数据类型添加到数据集中，供数据挖掘服务读取。 5、选择模型开发方式，可以是noteBook编程模式和可视化拖拽方式。我们选择第二种，即从“工作流管理”->“新建工作流” 在工作流列表页面点击工作流名称可以进入工作流编辑页面。 工作流编辑页面分为左中右三个区域。 左侧为算子区，显示当前工作流支持的所有算子。 中间为画布区，用于构建数据挖掘工作流。 右侧为属性区，用于显示算子或工作流的属性。



点击右侧的工作流配置，可以设置工作流的属性，例如工作流使用的资源等。

6、在工作流编辑界面完成工作流编辑后，运行生成模版，生成模型。

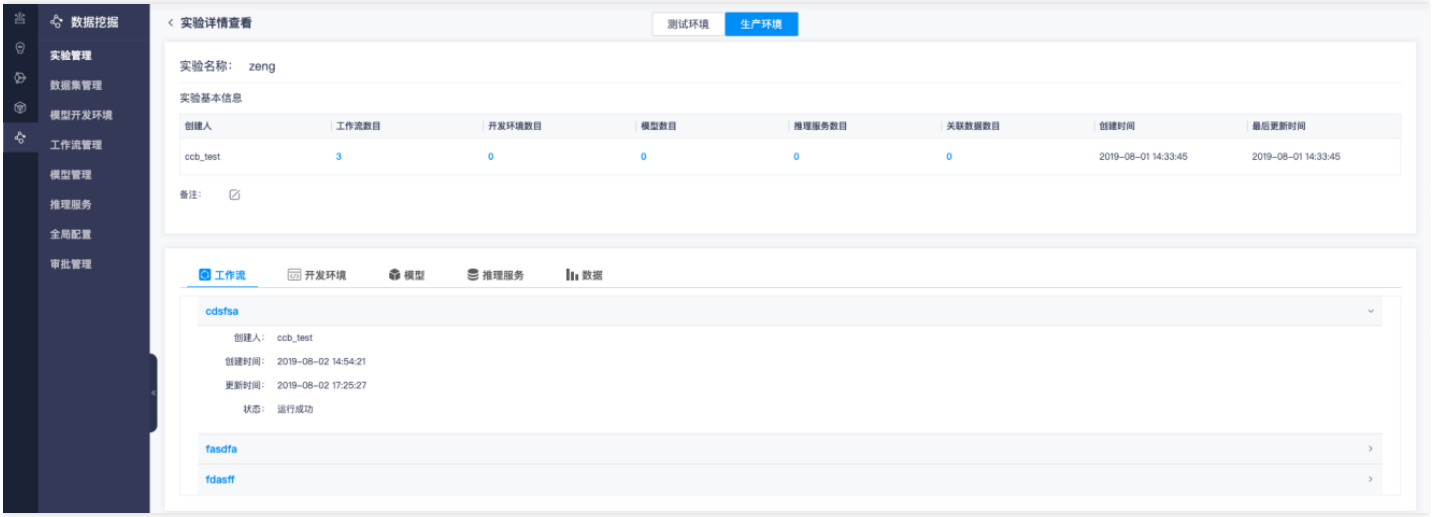
操作指南
实验管理

最近更新时间: 2019-11-13 07:08:31

点击大数据开发套件，点击项目空间，选择一个开通数据挖掘的项目进入数据挖掘模块，点击【实验管理】-新建实验，输入实验名称，选择实验类型，填写实验备注，点击【确认】按钮。

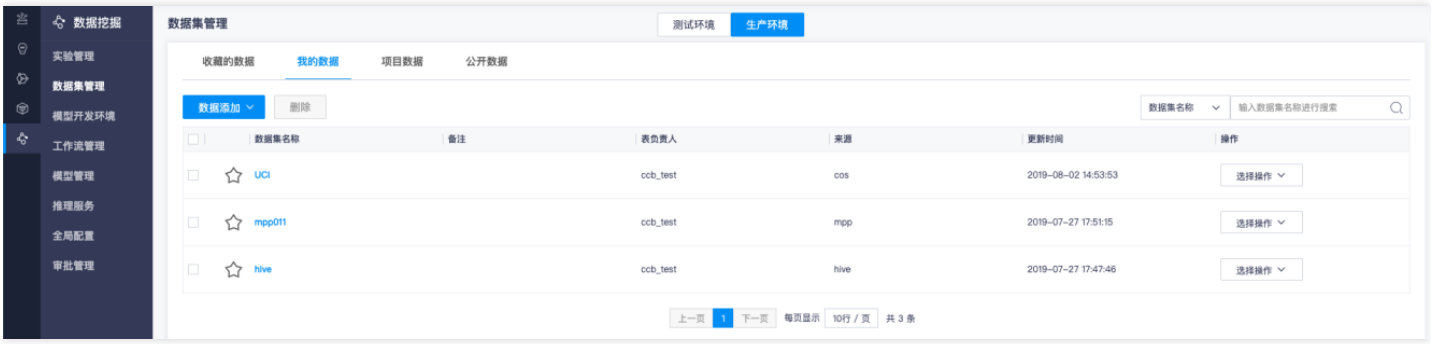
在实验列表展示一条新建的实验。

在实验详情页可以显示实验管理的资源，点击具体的关联资源，可以直接跳转到对应的资源页面。



数据集管理

最近更新时间: 2019-11-13 07:07:38

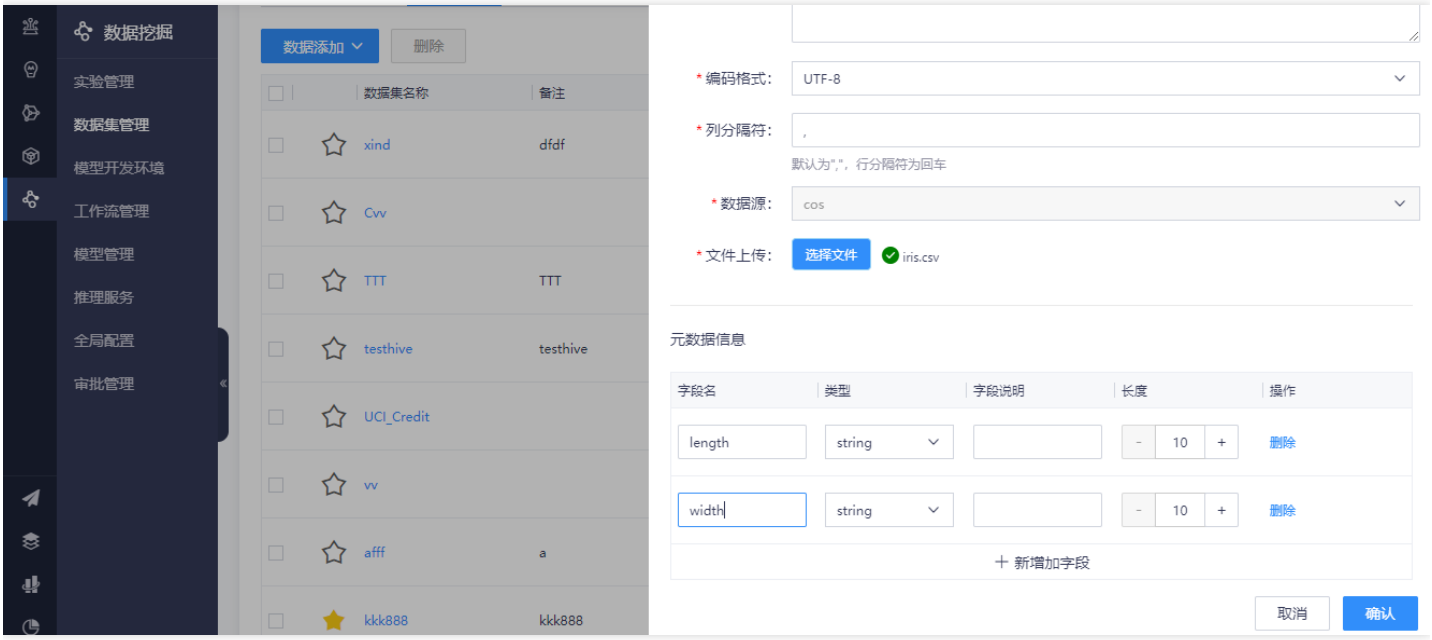


数据挖掘组件中有四类数据集： 我的数据：当前用户有权限的数据集。 项目数据：当前项目有权限的数据集。 公开数据：当前租户设置为公开的数据表。 收藏的数据：来自于我的数据集、项目的数据集的收藏。

数据集添加

最近更新时间: 2019-11-13 07:07:38

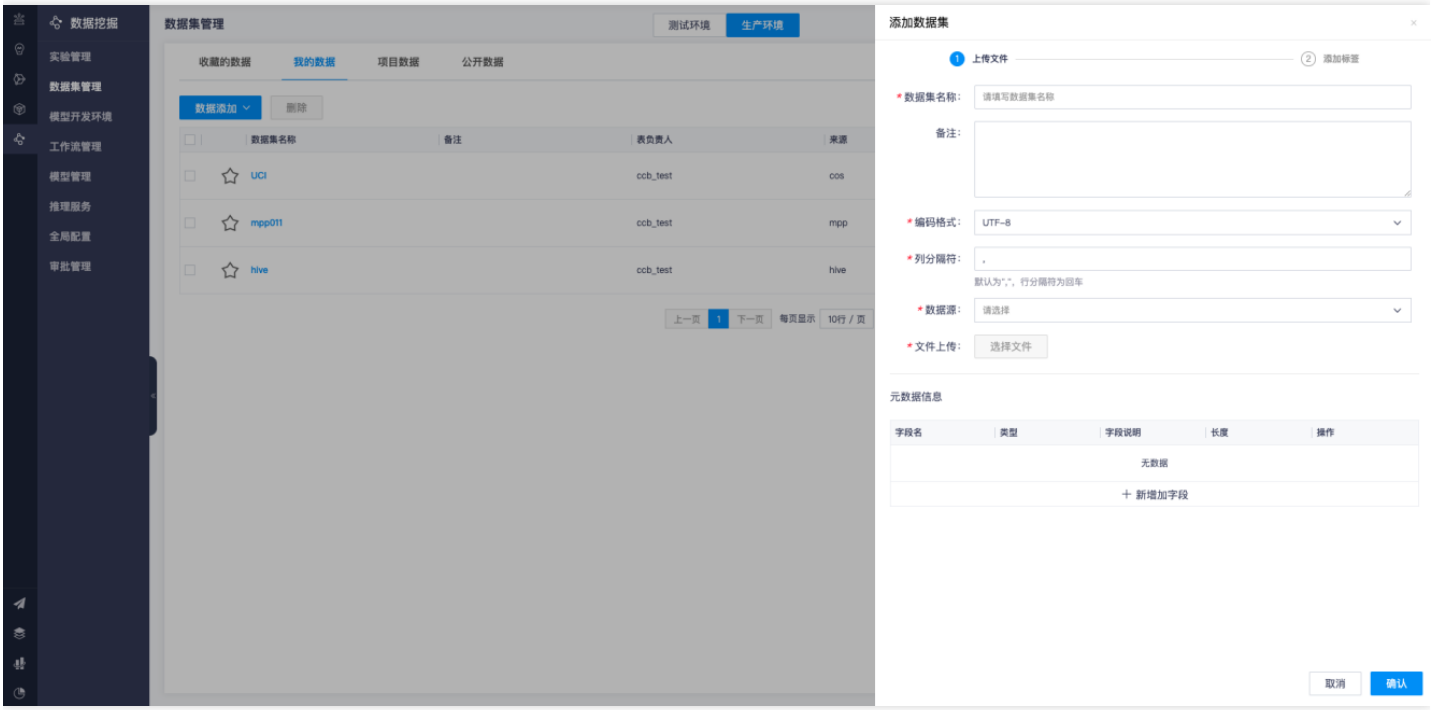
我的数据有两种数据集添加方式： 从本地上传文件到cos 从本地上传数据，需要指定数据集名称、数据文件编码方式及列分隔符。同时需要指定上传的数据源，以及数据的元数据信息。其中元数据信息中需要指定字段名和数据类型等信息。



上传文件后还可以添加对应的数据标签。设置相应的一级标签和二级标签。

从数据管理已有的数据表添加

从数据管理添加，可以将数据管理已有的数据类型添加到数据集中，供数据挖掘组件读取。



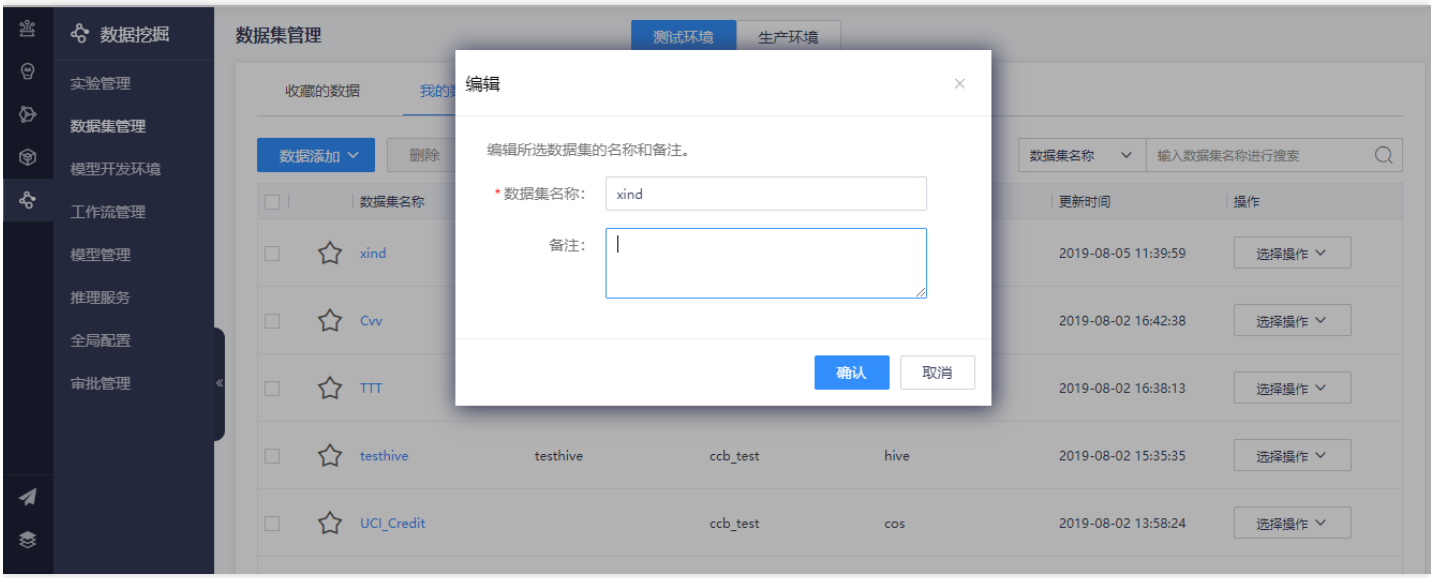
数据集要求：数据集名称命名规范，如若数据集内容为图片格式，应该保证图片质量，不能有损坏的图片，每一类数据若用户选择从本地上传到cos方式，支持上传csv格式文件，文件大小不超过1GB。

数据集分享

最近更新时间: 2019-11-13 07:07:38

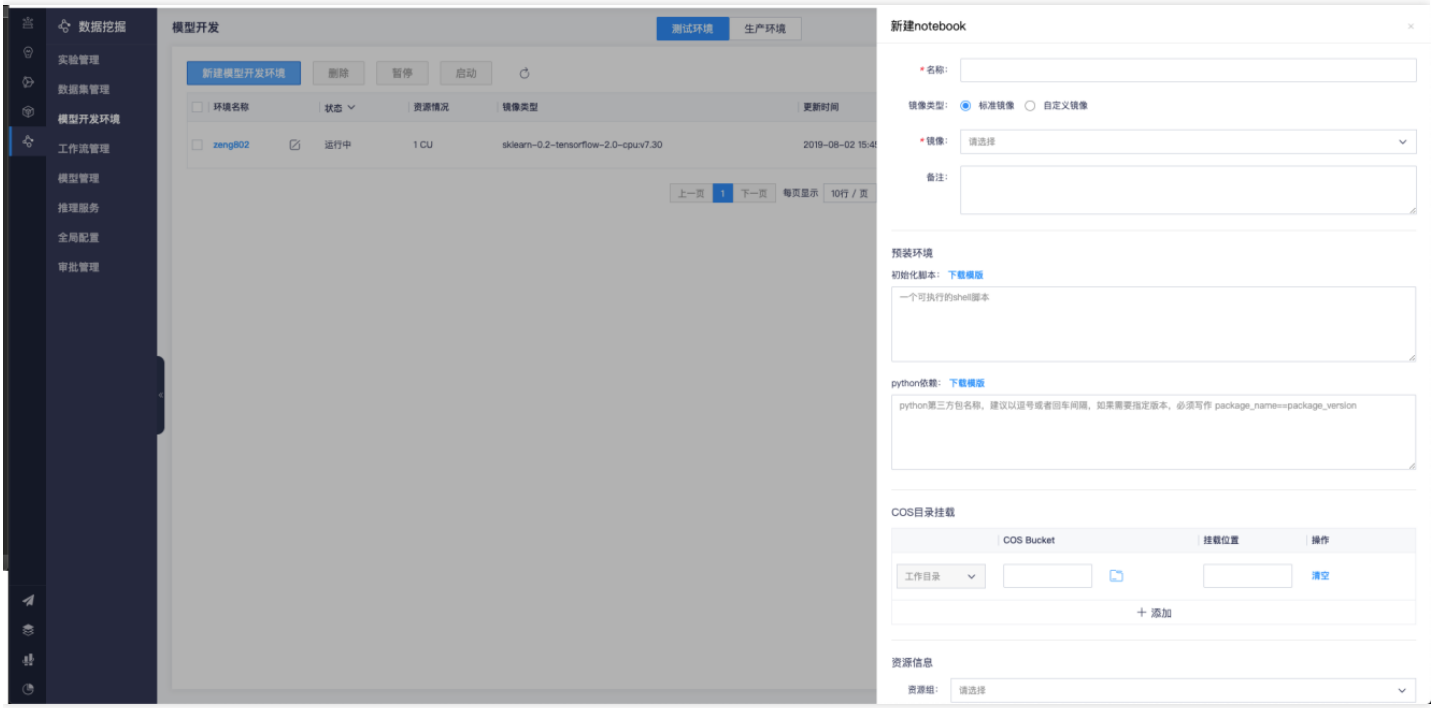
数据集可以共享给个人或项目，点击操作中的【共享】按钮，选择要共享的个人或项目。

您可以对即将共享的数据集添加备注，如数据集内容，标签信息等详情。



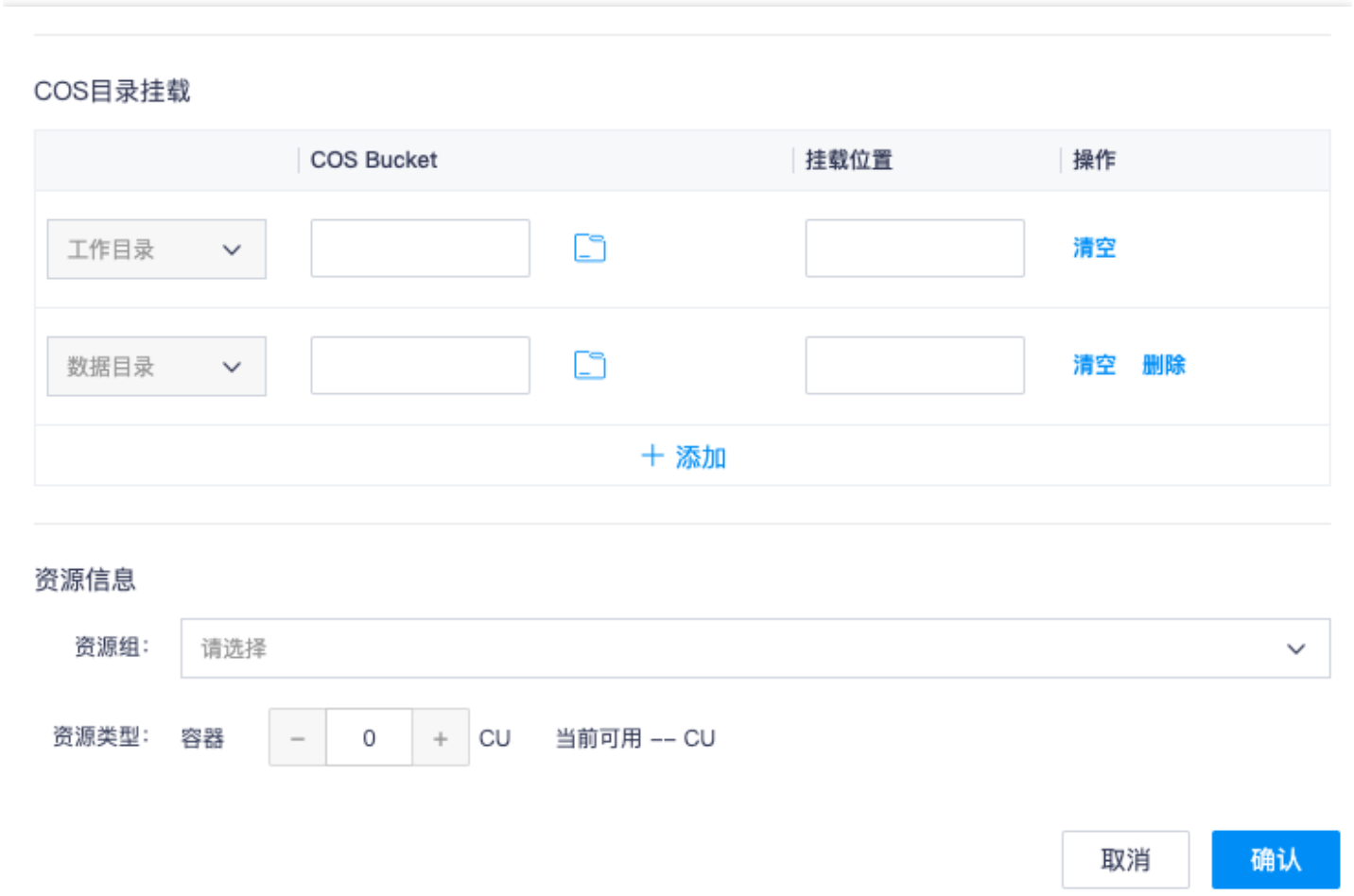
模型开发环境

模型开发环境是用于开发数据挖掘模型的环境。 新建模型开发环境



指定开发环境名称，选择镜像类型。支持两类镜像，一类是系统默认的标准镜像，一类是用户自定义的镜像。

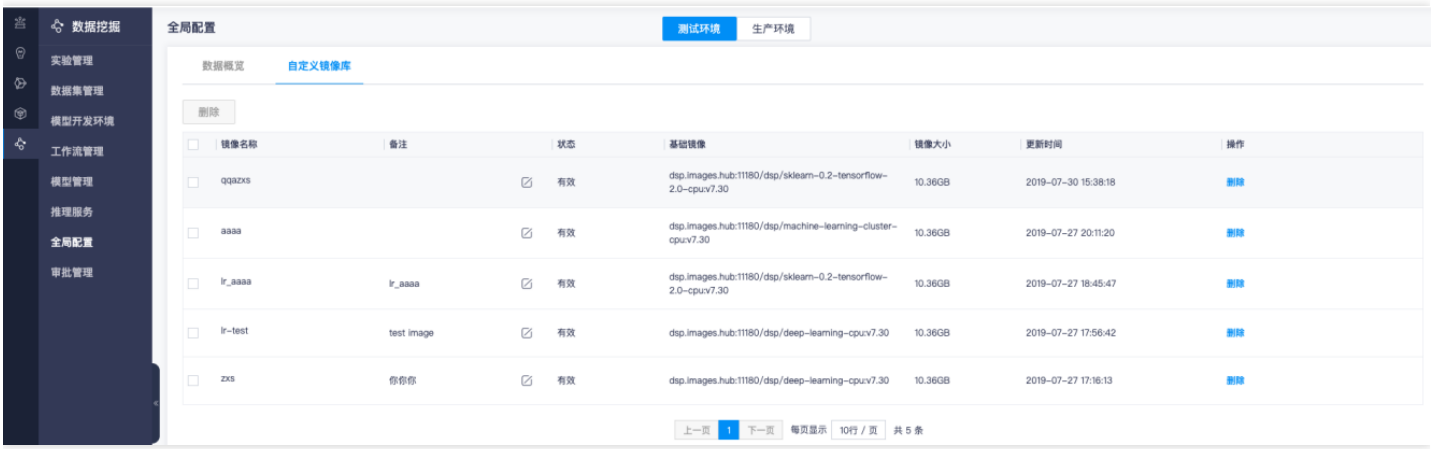
用户可以通过预装环境，对镜像环境进行自定义，数据挖掘组件提供两种方式对环境进行自定义： 通过初始化脚本，用户可以提供一个可执行的shell脚本，对环境进行定制 Python依赖，用户可以提供一个piprequirements.txt文件，来安装自定义的pip包



开发环境可以挂载cos目录作为工作目录和数据目录。 开发环境启动时需要指定需要使用的资源类型。需要指定资源组（参见项目管理）及对应的计算资源CU。 模型开发列表操作

模型开发列表中提供了各种针对模型开发环境的操作： 删除：删除对应开发环境 暂停/启动：暂停或启动模型开发环境 生成自定义镜像：将当前开发环境生成自定义镜像，后续可以用对应自定义镜像直接创建开发环境

生成的自定义镜像会出现在全局配置中，可以在其中进行备注或删除。 Lab模型打开/notebook模型打开：数据挖掘组件提供了两种jupyter打开的方式，lab和notebook。Jupyter Notebook功能强大，支持多种kernel，如Python3，PySpark、R和Spark等，相应的kernel配置方式见附录。

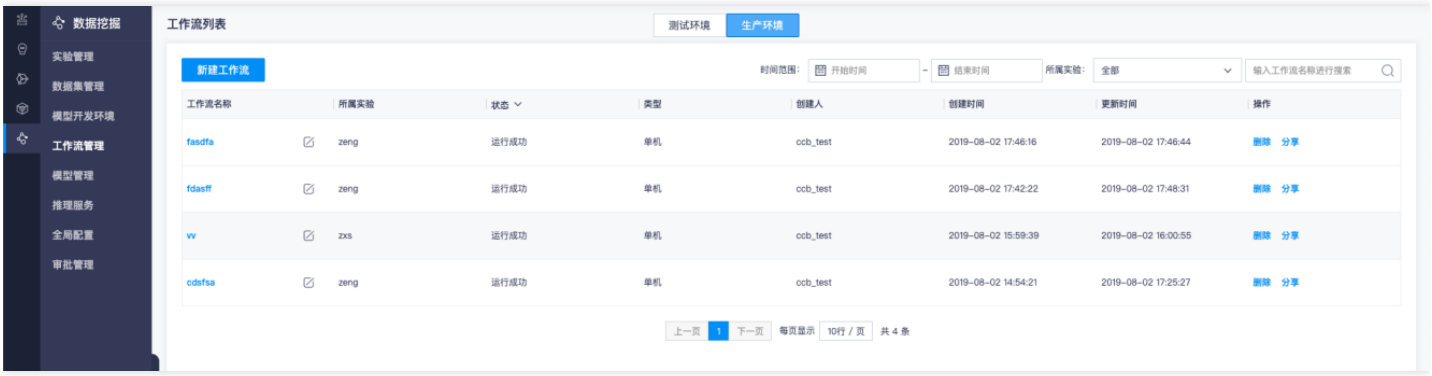


模型开发环境详情页 点击模型开发列表中的模型开发环境名称，可以进入模型开发详情页。

模型开发环境详情页显示开发环境的基本信息，COS挂载数据，以及实时的资源使用情况。

工作流管理

最近更新时间: 2019-11-15 07:10:03

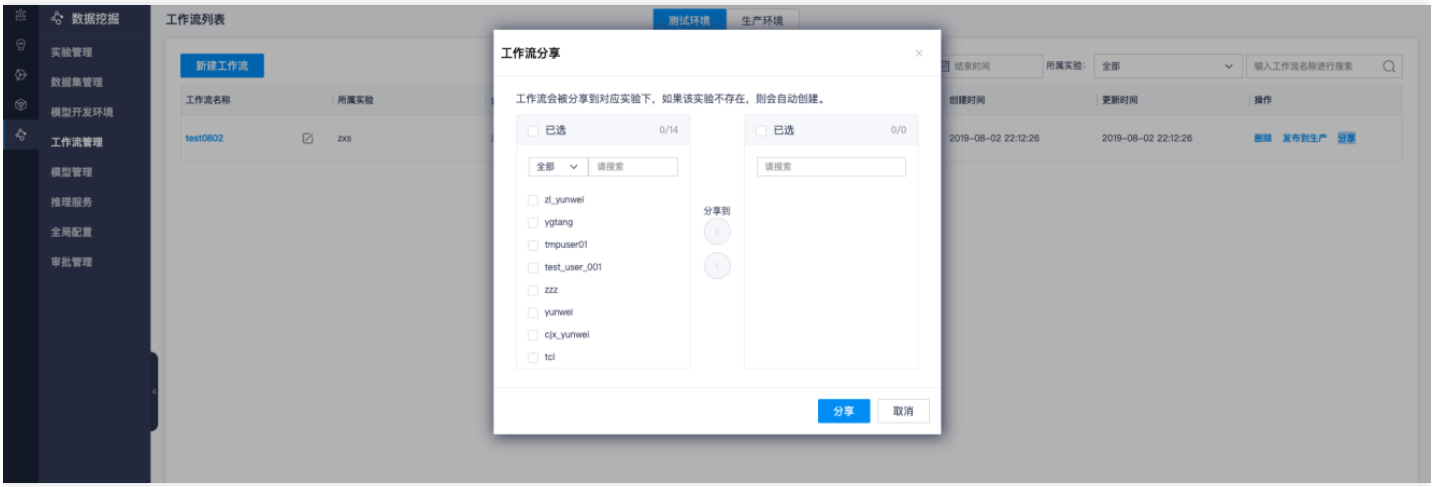


工作流是数据挖掘组件提供的可视化模型开发方式，用户可以通过拖拽算子的方式来构建模型训练过程。工作流列表页面显示现有的工作流的信息。 新建工作流

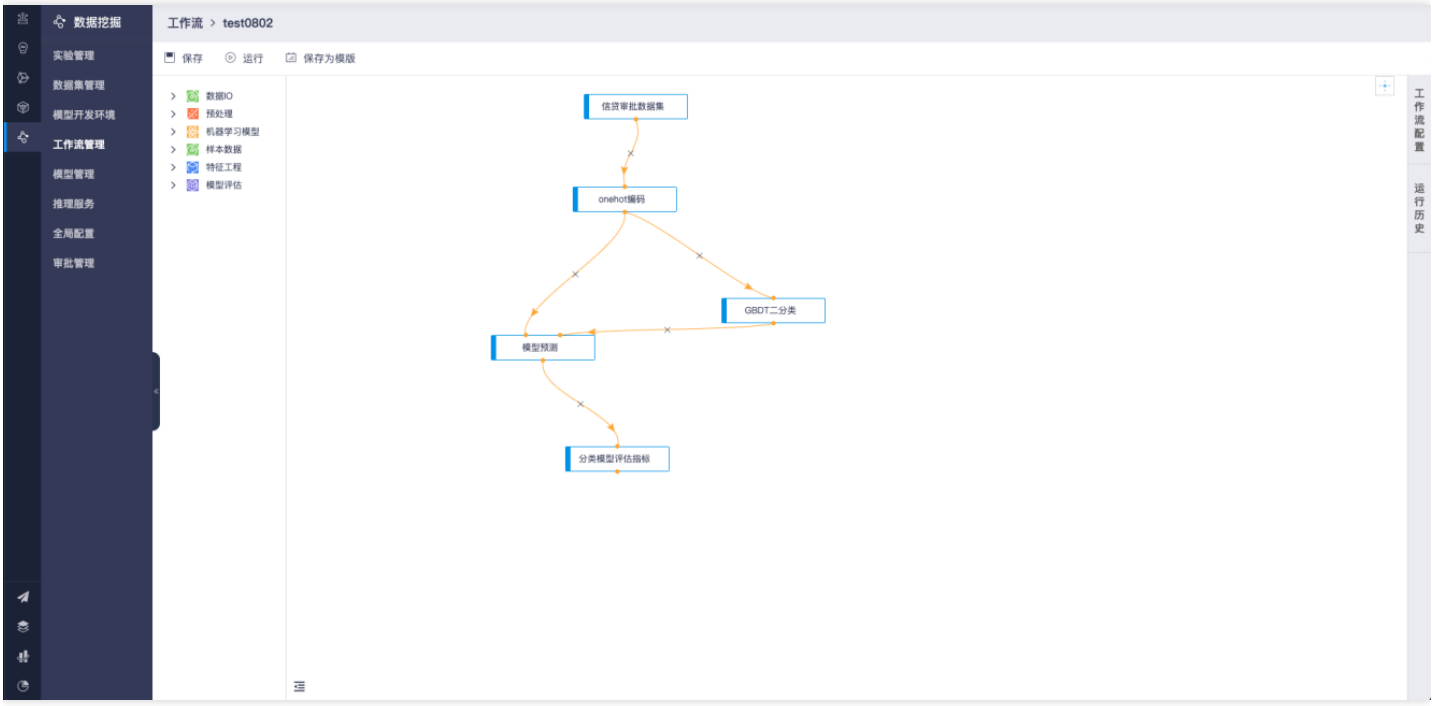
点击【新建工作流】按钮进入工作流创建页面，新建工作流需要指定工作流名称，所属实验，工作流类型以及所用的模板。 工作流根据底层资源的不同分为单机和集群版本。单机版提供sklearn、lightGBM、XGBoost等算法框架封装的算子，集群版提供SparkMLlib封装的算子。 工作流模板支持平台提供的公共模板与个人保存的个人模板。

新建完成的工作流会出现的工作流列表中，工作流列表支持的操作包括： 删除：删除对应工作流 发布到生产（仅测试环境有效）：将工作流从测试环境发布到生产环境。

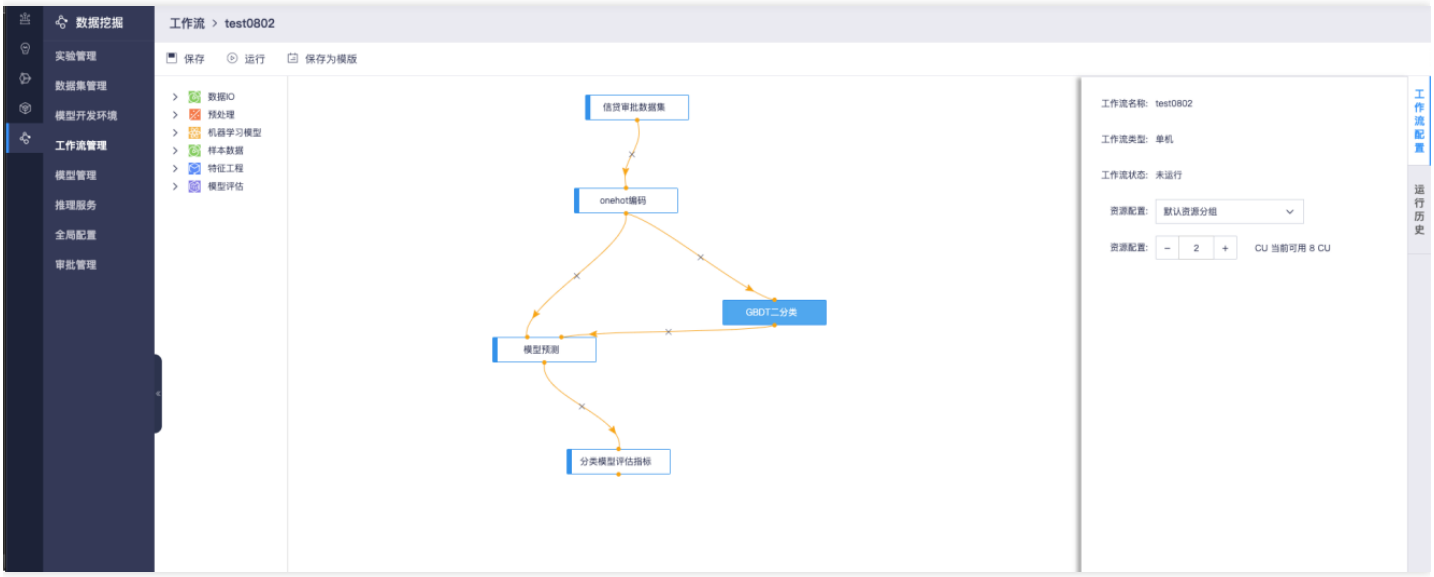
点击发布到生产，需要填写备注信息，需要由项目管理员审批，相关审批请求需要进入审批管理界面查看。 分享：将指定工作流分享给指定用户



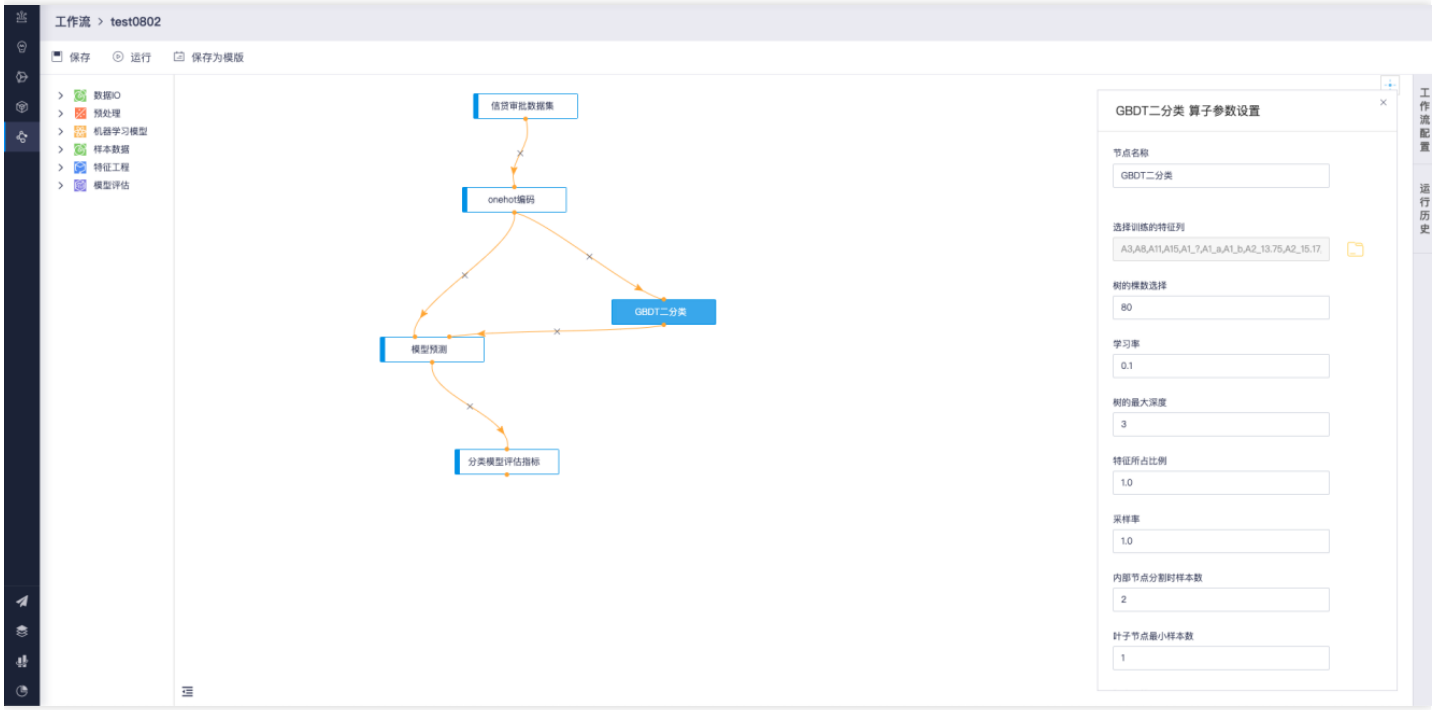
编辑工作流



在工作流列表页面点击工作流名称可以进入工作流编辑页面。 工作流编辑页面分为左中右三个区域。 左侧为算子区，显示当前工作流支持的所有算子。 中间为画布区，用于构建数据挖掘工作流。 右侧为属性区，用于显示算子或工作流的属性。 点击右侧的工作流配置，可以设置工作流的属性，例如工作流使用的资源等。

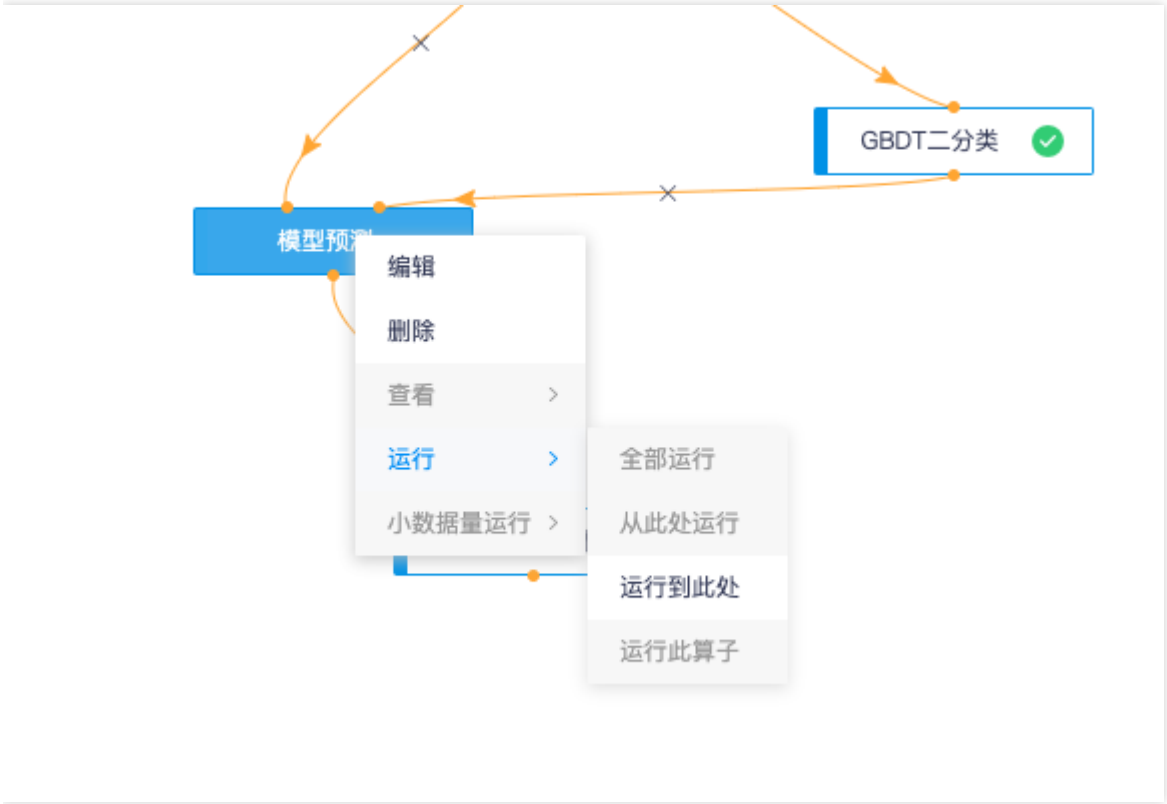


双击算子，右侧会显示算子的属性。根据算子的不同，算子可以配置的属性也不同，例如对于模型类算子GBDT二分类的算子主要是算法的超参数设置。



工作流顶部是工作流的操作区域，可以对工作流进行保存、运行以及将工作流保存为个人模板。

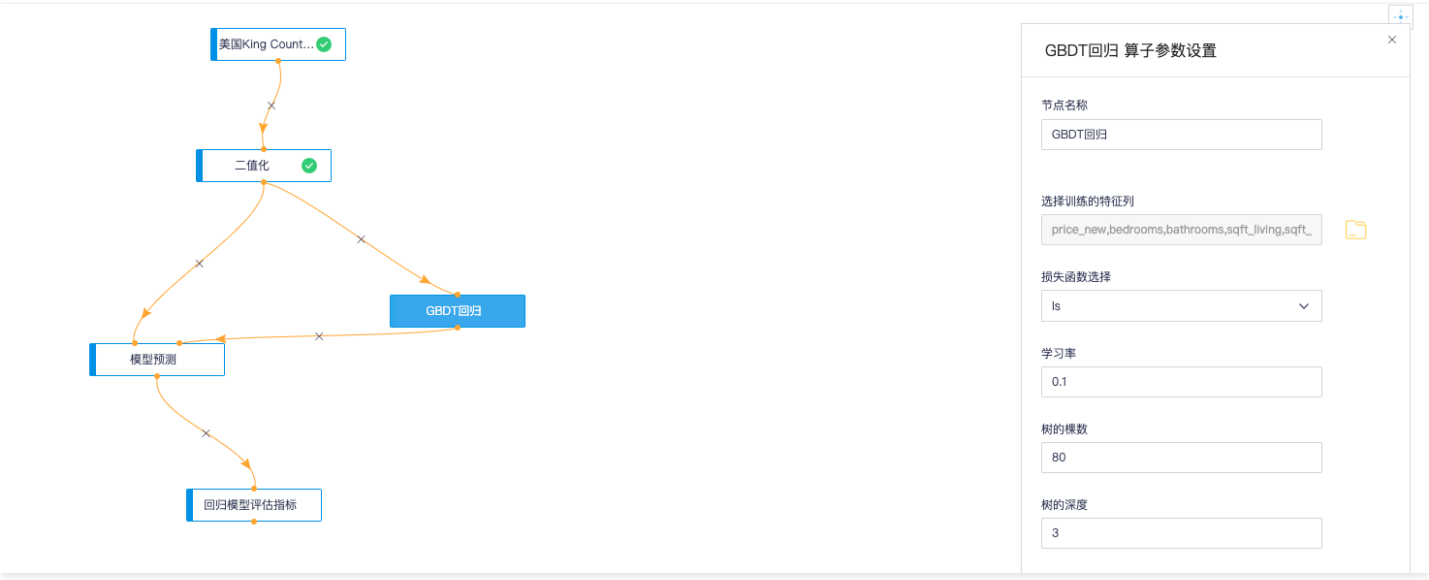
算子上点击右键，可以对算子进行更多的操作，不同的算子支持的右键操作略有差别。



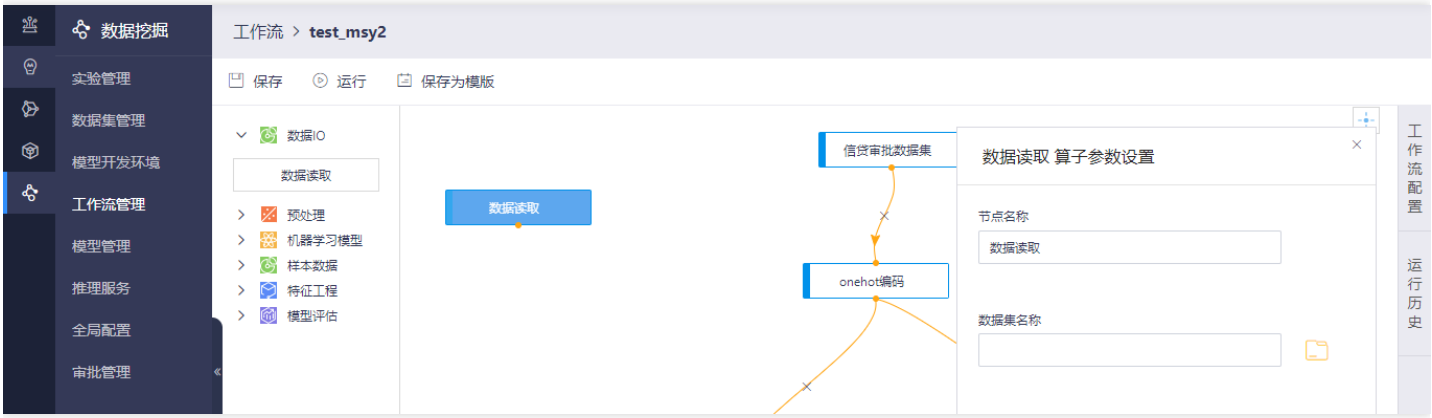
编辑：编辑算子参数； 删除：删除当前算子； 运行：可以指定运行完整工作流、运行到此处或者运行此算子。 算子和算子之间可以通过连线进行连接，通过点x符号可以删除算子之间的连线。

运行工作流 工作流运行需要指定运行工作流所需的资源CU/DCU，才能进行运行。工作流的资源配置在右侧的工作

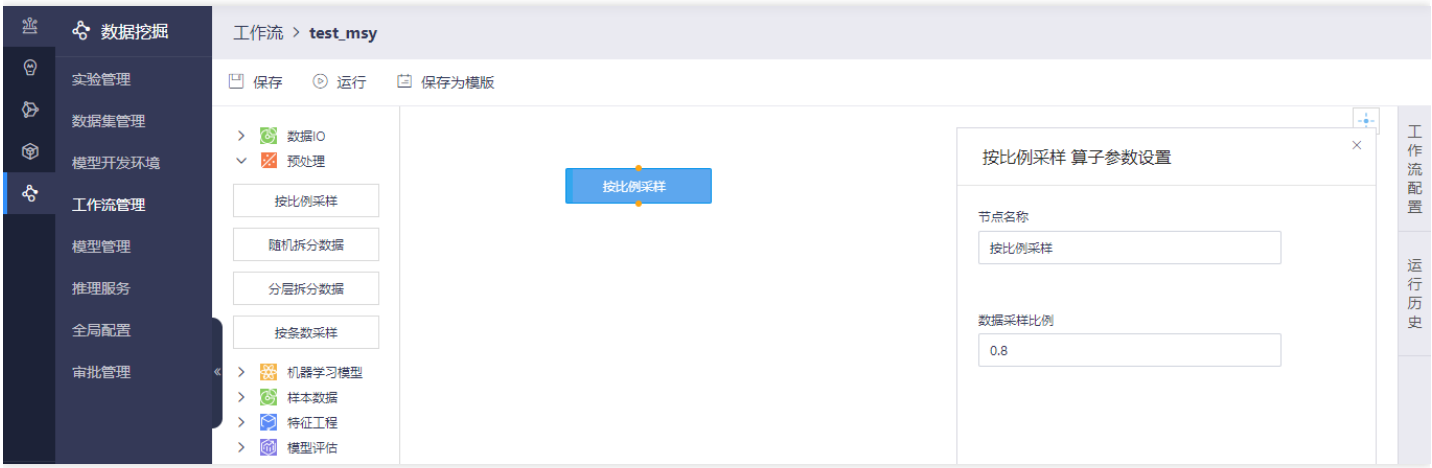
流配置中指定。 工作流的运行有两种方式，一种是通过右上的全部运行。另一种是通过算子上右键来进行运行。



模型类算子需要指定特征列与标签列，需要先运行模型类算子以上的算子，才能获得对应的特征列和标签列。 算子简介 以信贷审批项目为说明案例，在左侧算子区中分为数据IO、预处理、机器学习模型、样本数据、特征工程和模型评估等6种算子类型。所有算子均具有右击和双击操作。 数据IO：包含数据读取算子，右击算子可以重命名、删除和复制算子，查看日志，运行算子。双击算子可以设置算子参数，如节点名称和数据集名称。算子支持版本为Spark单机和集群版。

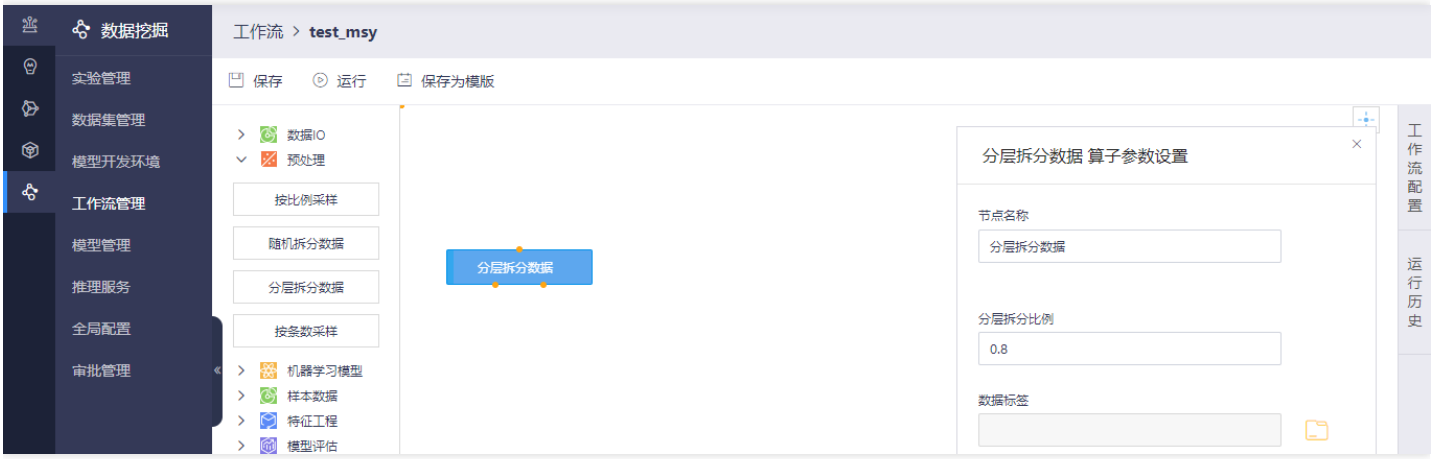


数据预处理：包括按比例采样、随机拆分数据、分层拆分数据和按条数采样4种算子类型。其中右击每个算子均支持重命名、删除、复制、运行算子操作，同时支持查看日志和支持小数据量运行操作，其中小数据量运行具体分为全部运行、运行到此和运行此算子操作。算子输入和输出数据格式均为DataFrame。 其中按比例采样算子页面支持设置节点参数操作，您可以输入数据集采样比例。

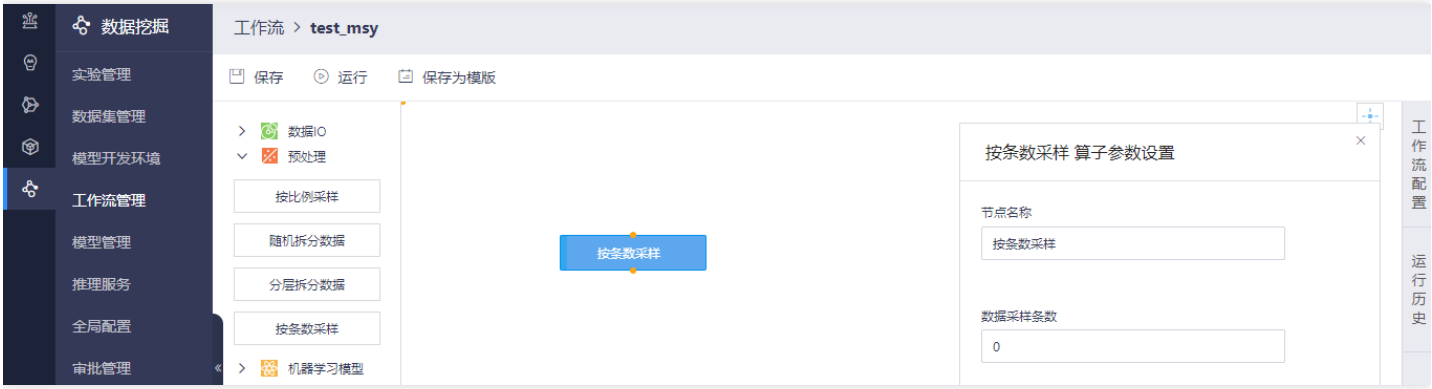


随机拆分数据算子页面支持设置算子参数操作，您可以输入数据拆分比例。

分层拆分数据算子页面支持设置算子参数操作，您可以输入分层拆分比例和数据标签等信息。



按条数采样算子页面支持设置算子参数操作，您可以输入数据采样条数。



机器学习模型：包含GBDT回归、xgboost二分类、K均值聚类和GBDT二分类四种算子类型。右击算子均支持重命名、删除、复制、运行算子操作，同时支持查看日志和支持小数据量运行操作，其中小数据量运行具体分为全部运行、运行到此和运行此算子操作。 其中GBDT回归算子支持解决回归任务，双击算子支持设置算子参数，参数涉及选择特征列、选择损失函数、学习率设置、数的颗数、树的深度、节点分割时方法、分割时的最小样本数、叶子节点最小样本数、采样率、最大特征所占比例和选择模型的标签。节点分割时的方法分为friedman_mse、mse和mae三种方法。

参数说明：

参数	说明	默认值
损失函数	度量模型输出的预测值，与实际值之间的差距的一种方式。损失函数分为ls、lad、huber和quantile四种。	ls函数
节点分割时的方法	分割节点，方法分为friedman_mse、mse和mae	friedman_mse

其中xgboost二分类算子支持解决二分类任务，双击算子支持设置算子参数，参数涉及选择训练的特征列、树的最大深度、学习率、树的数目、选择基学习器、叶子节点划分所需最小损失、样本采样率、特征采样率、最小叶子节点样本权重、L1正则化权重、L2正则化权重和选择的模型标签。 其中K均值聚类算子支持解决无类别标签任务，双击算子支持设置算子参数，参数涉及选择训练的特征列、聚类簇个数选择、初始中心点选择方法、迭代次数和误差收敛值精确度。 参数说明：

参数	说明	默认值
初始中心点选择	选择初始的中心点方法，支持k-means++和random方法	k-means++

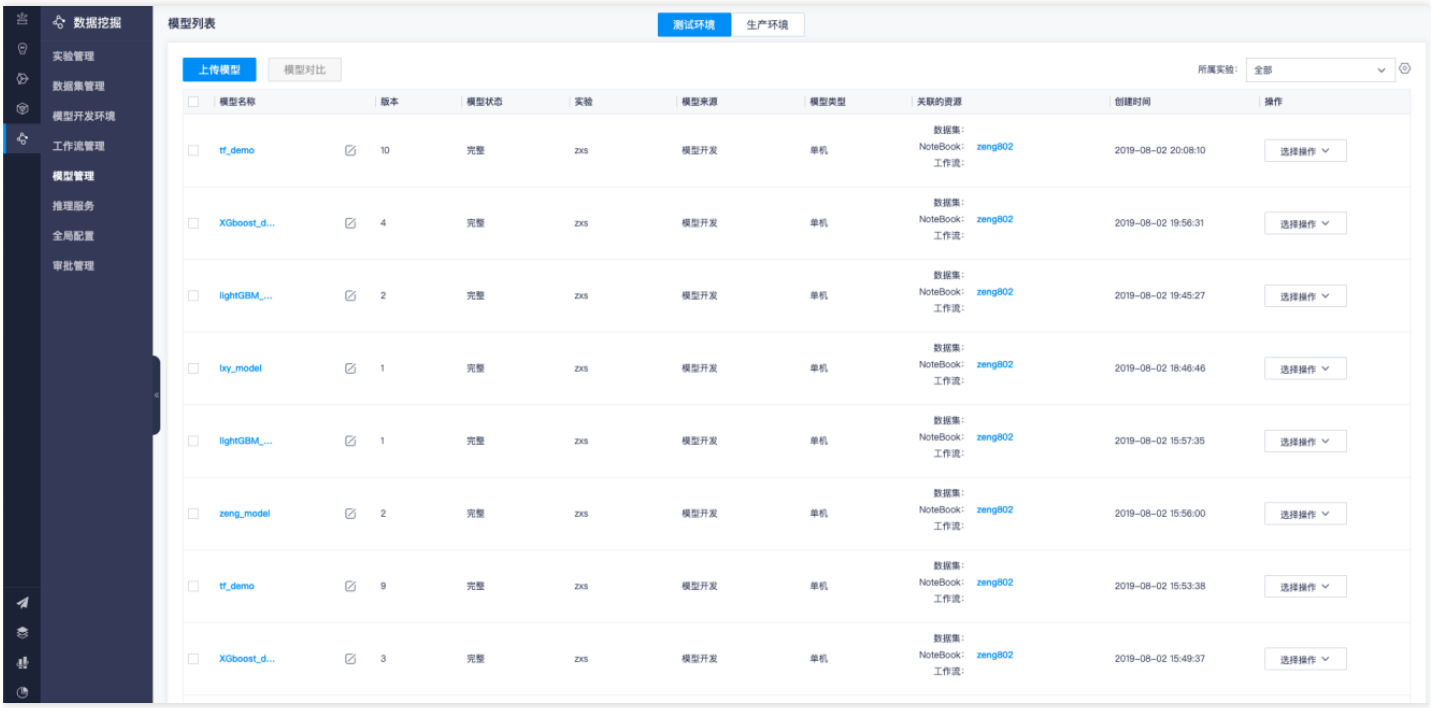
其中GBDT二分类算子支持解决二分类标签任务，双击算子支持设置算子参数，参数涉及选择训练的特征列、树的颗数选择、学习率、树的最大深度、特征所占比例、采样率、内部节点分割时样本数、叶子节点最小样本数、损失函数、分割时方法选择和选择模型标签。 参数说明：

参数	说明	默认值
损失函数	度量模型输出的预测值，与实际值之间的差距的一种方式。支持deviance和exponential	deviance
分割时方法选择	类别分割时的方法，支持friedman_mse、mse、mae	friedman_mse

样本数据：包含信贷审批、信用卡欺诈检测、电信客户流失数据、美国King County的房屋销售数据、乳腺癌、商城客户细分数据、台湾信用卡借贷等数据集。右击算子均支持重命名、删除、复制、运行算子操作，同时支持查看日志和支持小数据量运行操作，其中小数据量运行具体分为全部运行、运行到此和运行此算子操作。 特征工程：分为最大最小归一化、标准归一化、二值化和onehot编码四种算子。右击算子均支持重命名、删除、复制、运行算子操作，同时支持查看日志和支持小数据量运行操作，其中小数据量运行具体分为全部运行、运行到此和运行此算子操作。算子支持版本为Spark单机和集群版。算子输入和输出数据格式均为DataFrame。 模型评估：分为分类模型评估指标，聚类模型评估指标、回归模型评估指标和模型预测。双击算子支持查看运行后的模型指标。 其中分类模型评估指标，聚类模型评估指标、回归模型评估指标中均表征模型的表现详情，具体为AUC值、混淆矩阵、召回率、准确率、精确率和F1 score指标。 其中模型预测算子支持选择特征列表和选择模型标签。

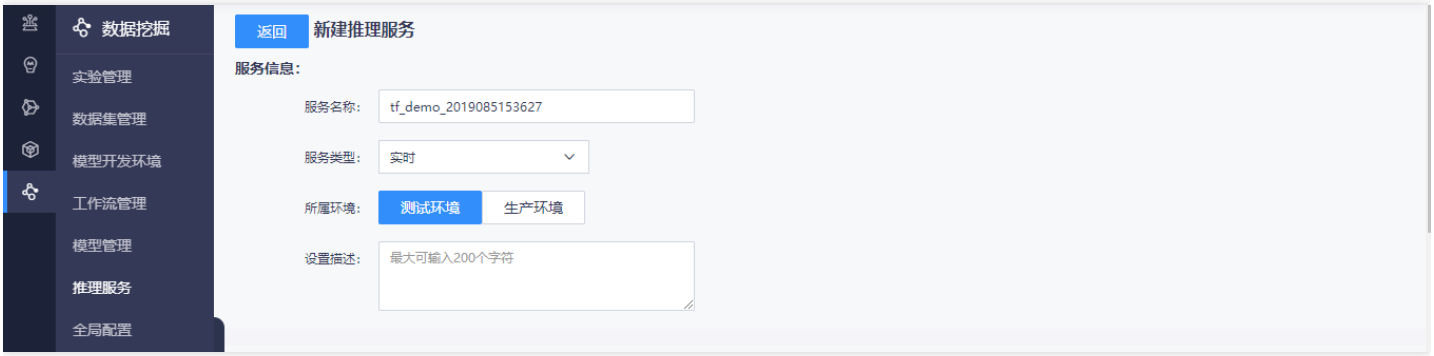
模型管理

最近更新时间: 2019-11-15 07:28:24

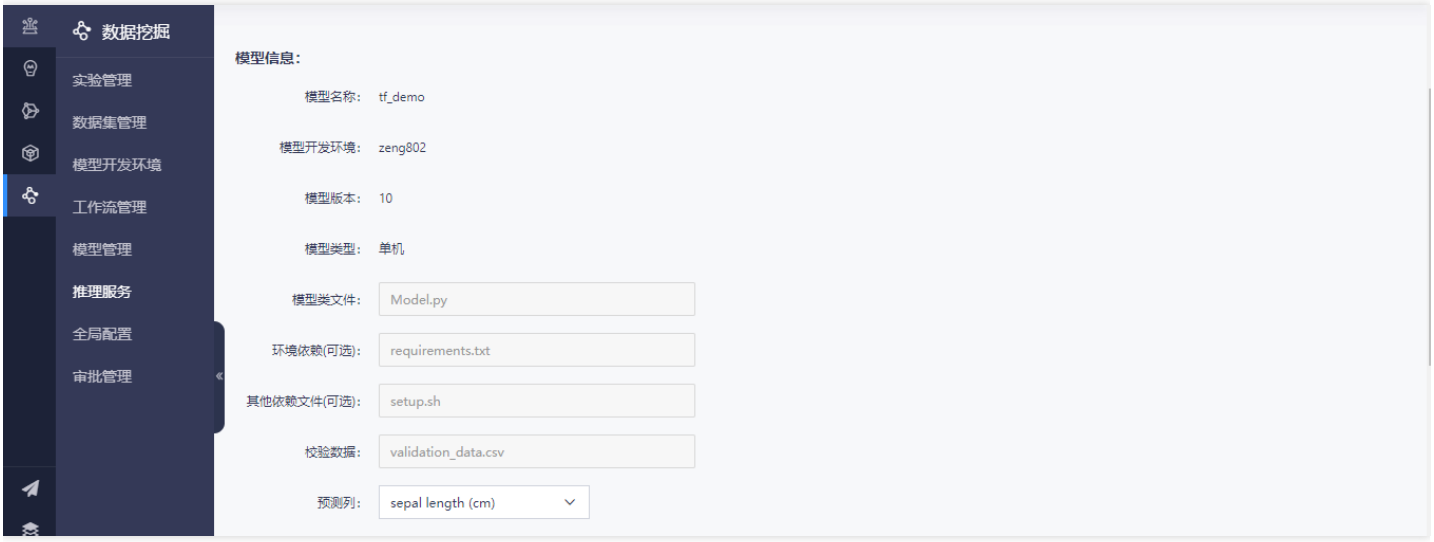


模型管理中显示在平台内生成的模型，显示模板的版本及相关资源等信息。

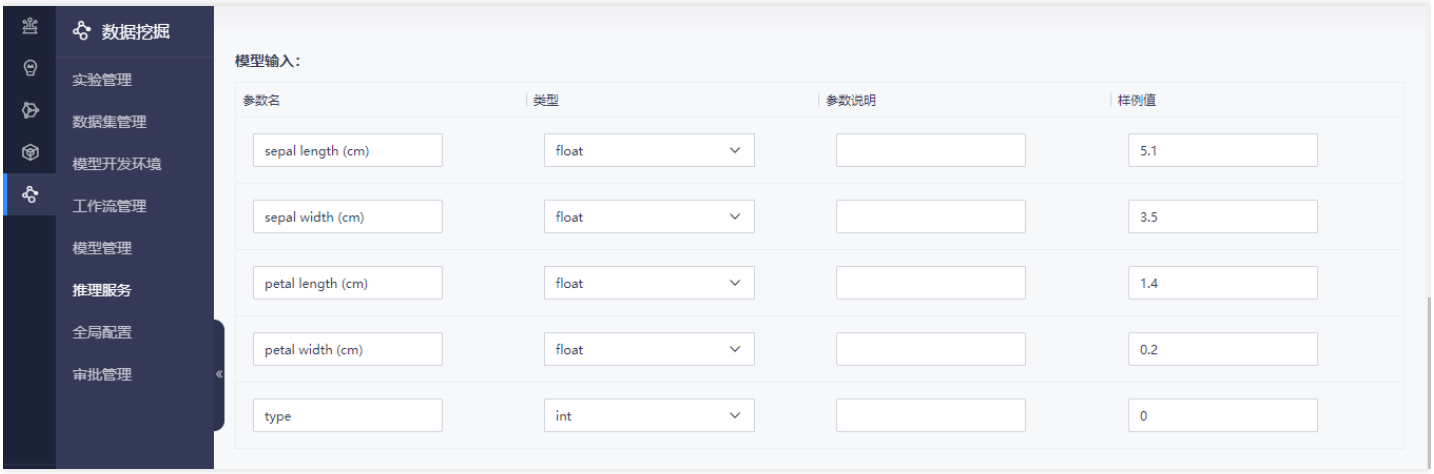
模型列表中支持的操作包括：删除、创建推理服务和发布到生产环境和分享操作。其中删除即为删除模型，创建推理服务即为从模型创建一个推理服务，发布到生产环境即为仅在测试环境，将测试环境的工作流发布到生产环境，需要项目管理员审批。 创建推理服务：点击创建推理服务按钮，进入创建推理服务页面，用户设置相应的服务信息、模型信息、模型输入和资源信息。其中在服务信息部分，用户可以设置服务类型和所属环境。



模型信息可以选择需要预测的属性列，同时显示模型开发环境、模型版本和校验数量等详细信息。



模型输入部分可以设置相应的参数名和数据类型，便于用户的使用，用户同时可以设置样例值。



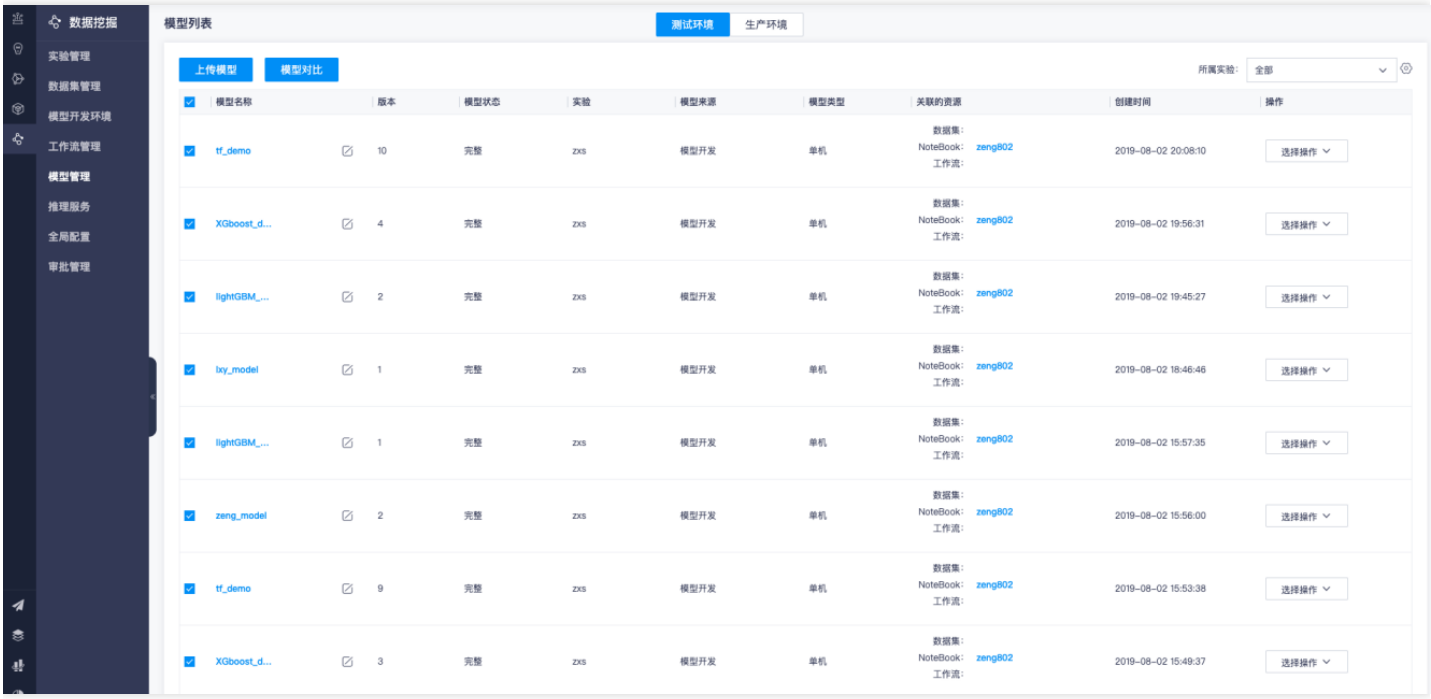
在资源信息部分用户可以选择资源组，资源类型和分配的资源个数。

创建推理服务成功后会在模型管理界面显示模型正在创建中等状态。

发布到生产环境：点击发布到生产环境，输入模型备注信息，点击发布，实现模型发布操作。

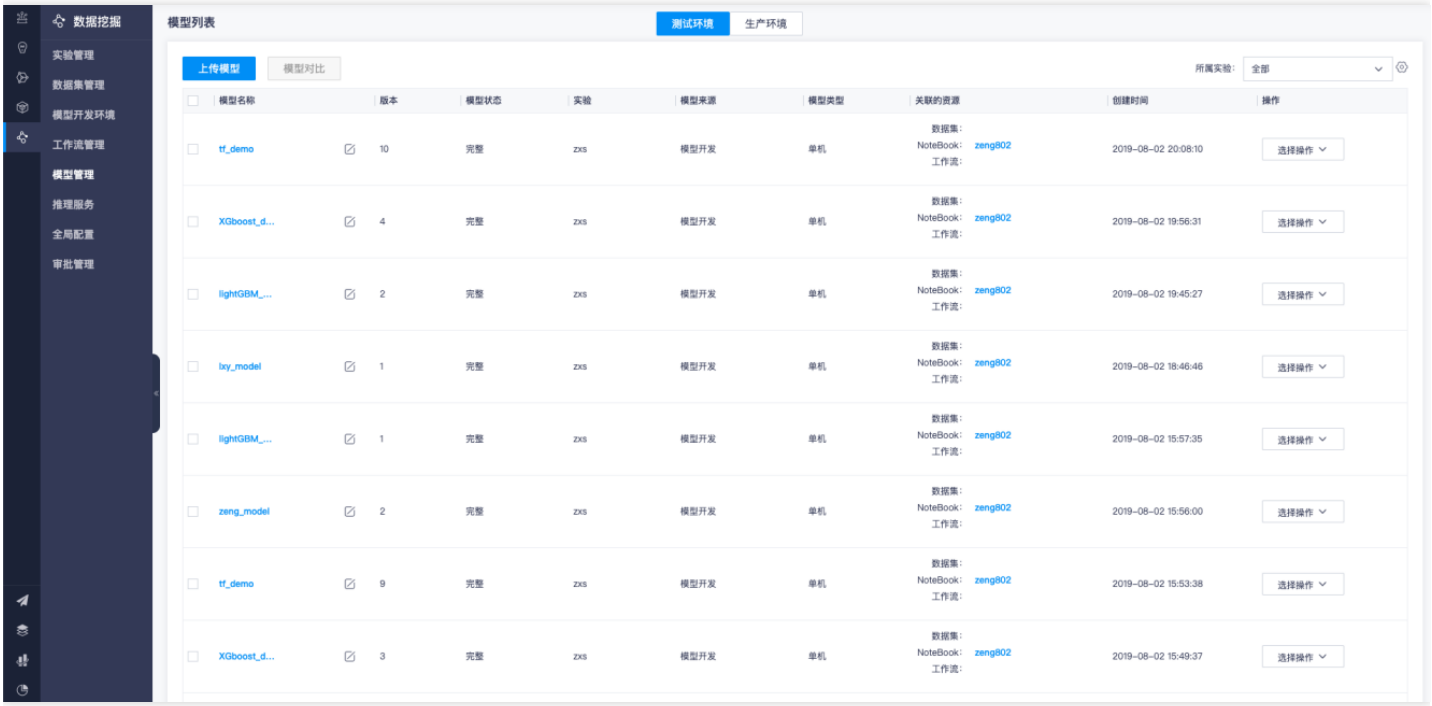
分享：将模型分享给其他用户，可以选择被分享对应的模型和分享对象。 删除模型：选择待删除的模型，点击删除实现删除模型操作。

模型对比

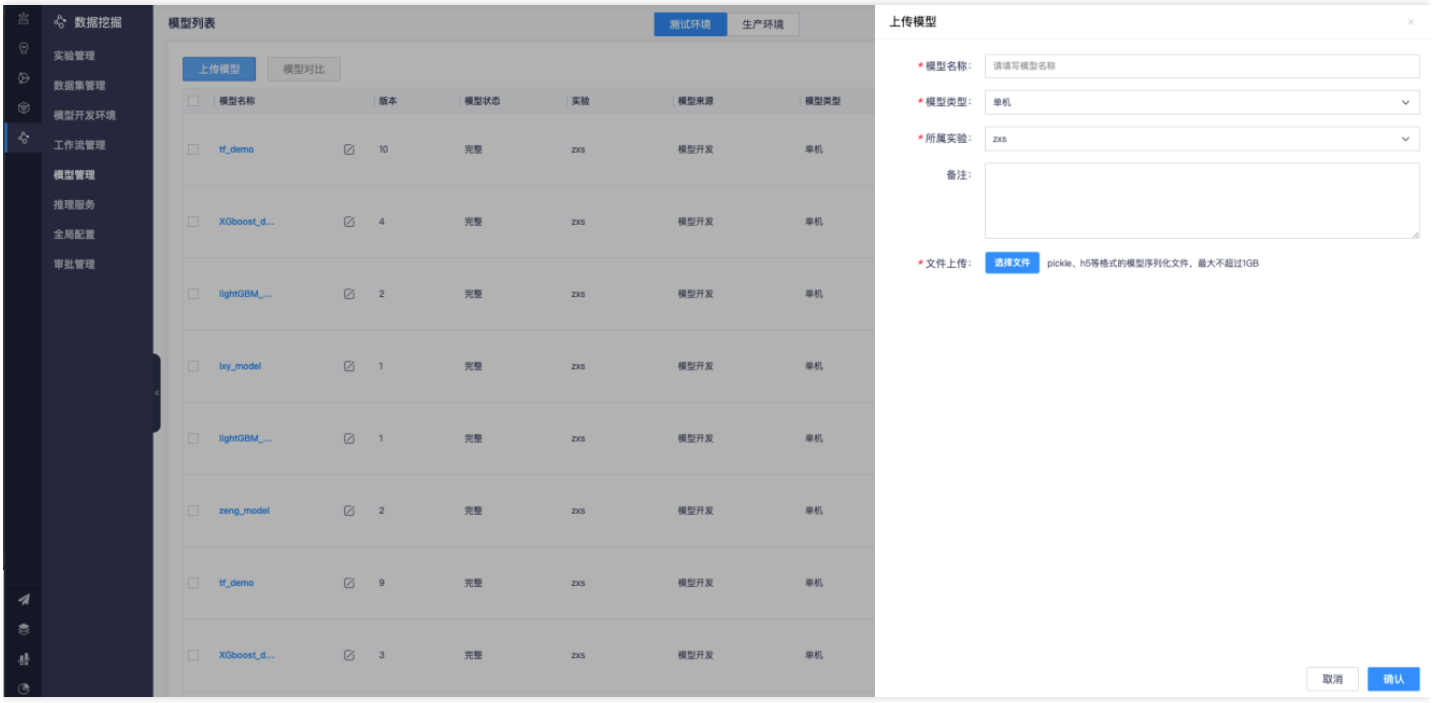


在模型列表中选择多个模型后，可以对多个模型的超参数及模型指标进行对比。

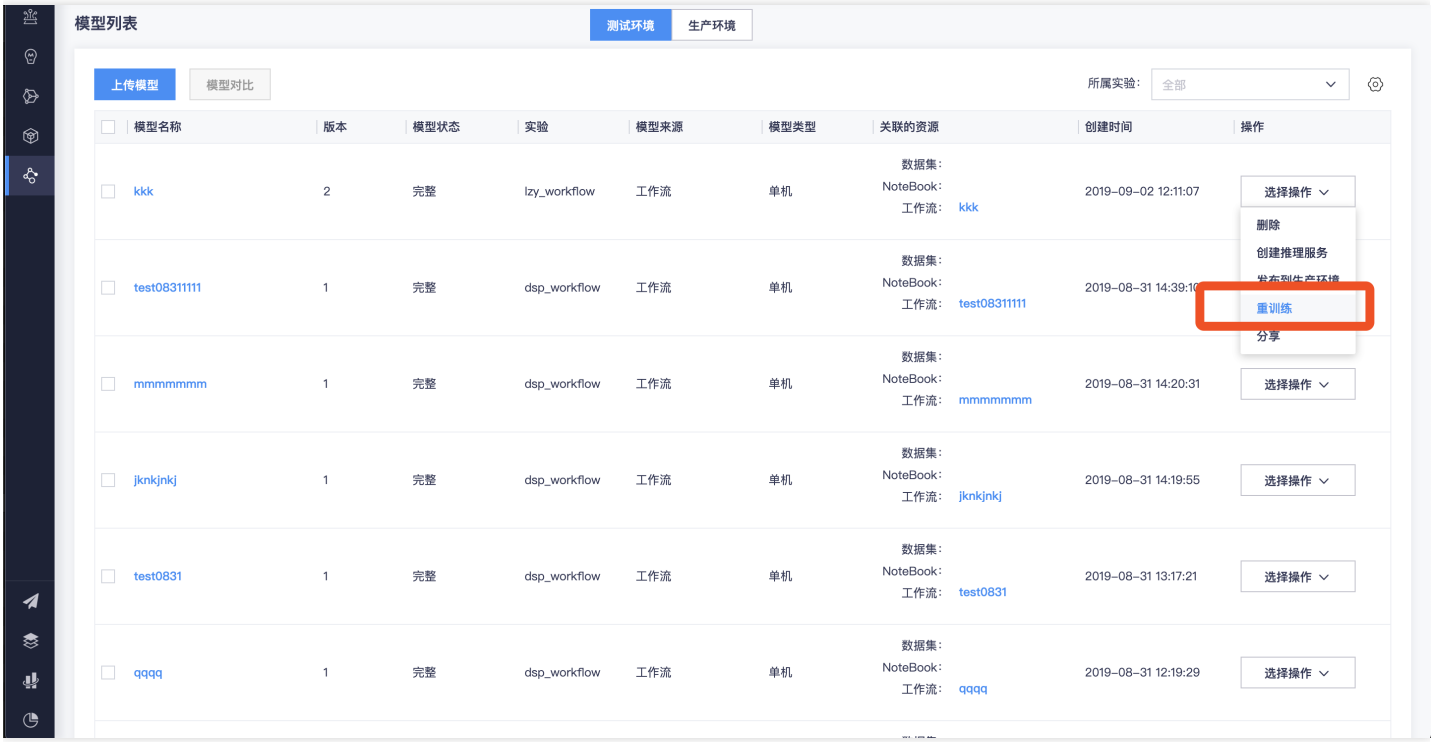
用户可以选择两个指标进行对比，了解不同超参以及指标间的共变关系。超参数名称具体分为max_depth、max_features、评估指标、auc、accuracy_rate、rmse和r2。 上传模型



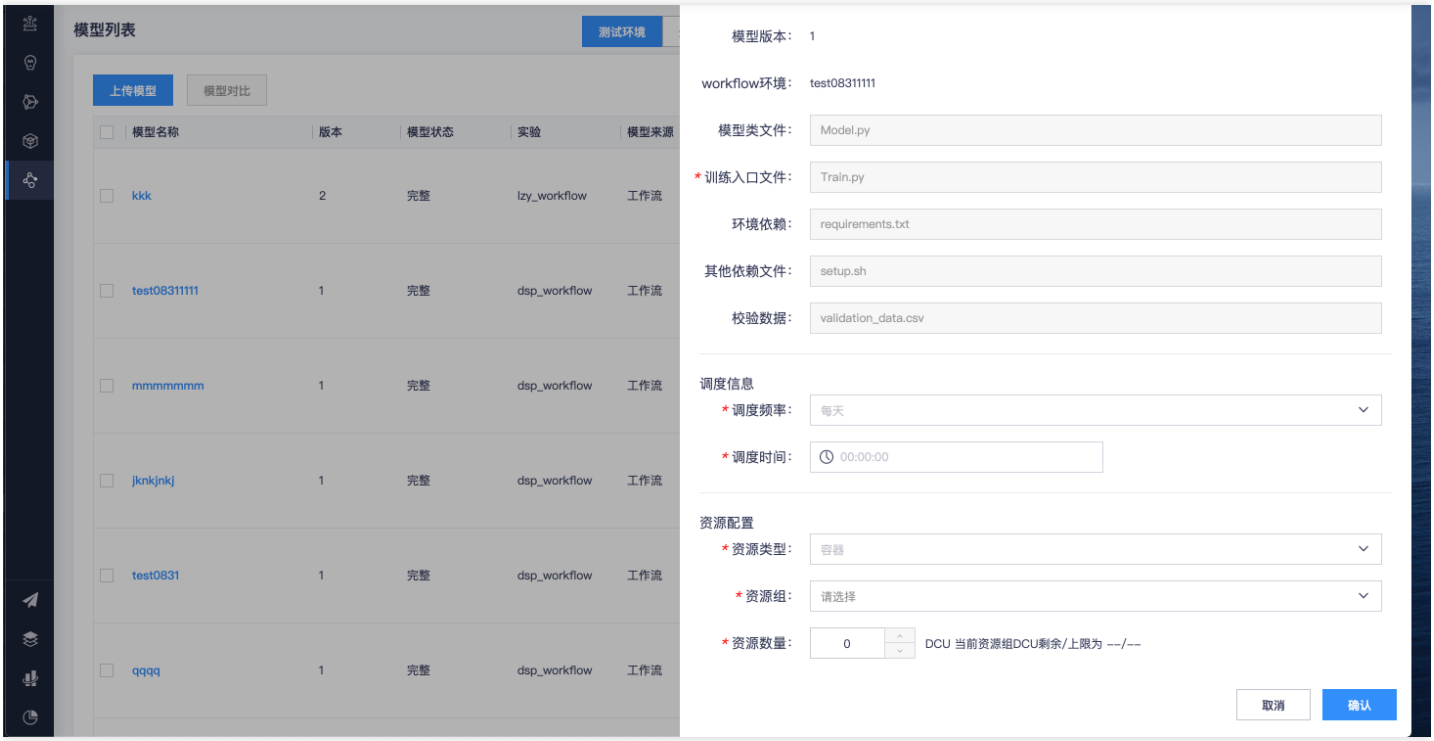
在模型列表页点击上传模型按钮，可以将平台外部训练的模型上传到平台的模型管理中。模型支持单机和集群两张类型，选择所属实验和点击上传文件，实现上传模型。



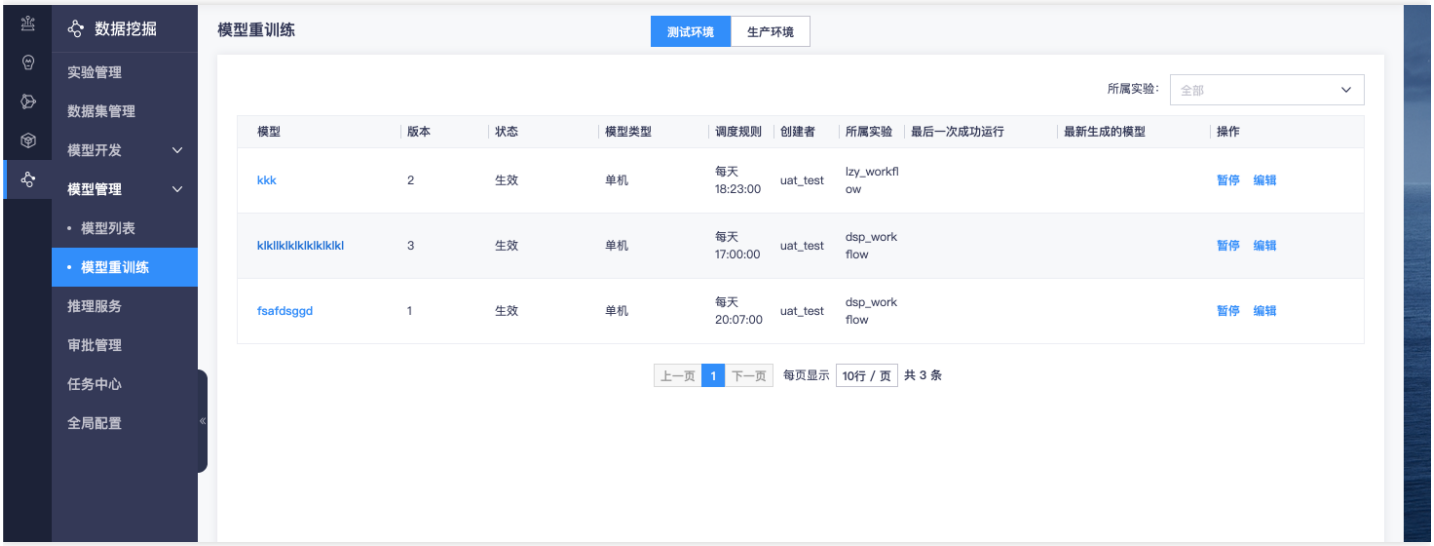
模型要求说明：模型支持pickle、h5等格式的模型序列文件，每个模型大小不大于1GB，模型名称遵循文件名规范。 模型重训练 由于户行为的变化或对模型算法的适应性，随着时间的变化，模型的效果会有所下降，因此，需要引入一定的机制刷新模型，保证线上模型的效果，在数据挖掘组件中，用户可以通过模型重训练功能来进行模型更新。



在模型列表中，每个模型的操作栏都带有重训练选项，用户可以通过这个选项来进行重训练的设置。



用户设置调度信息、资源信息来设置重训练任务。调度信息用于设置触发调度的周期规则，资源信息用于指定当前重训练任务使用的资源。

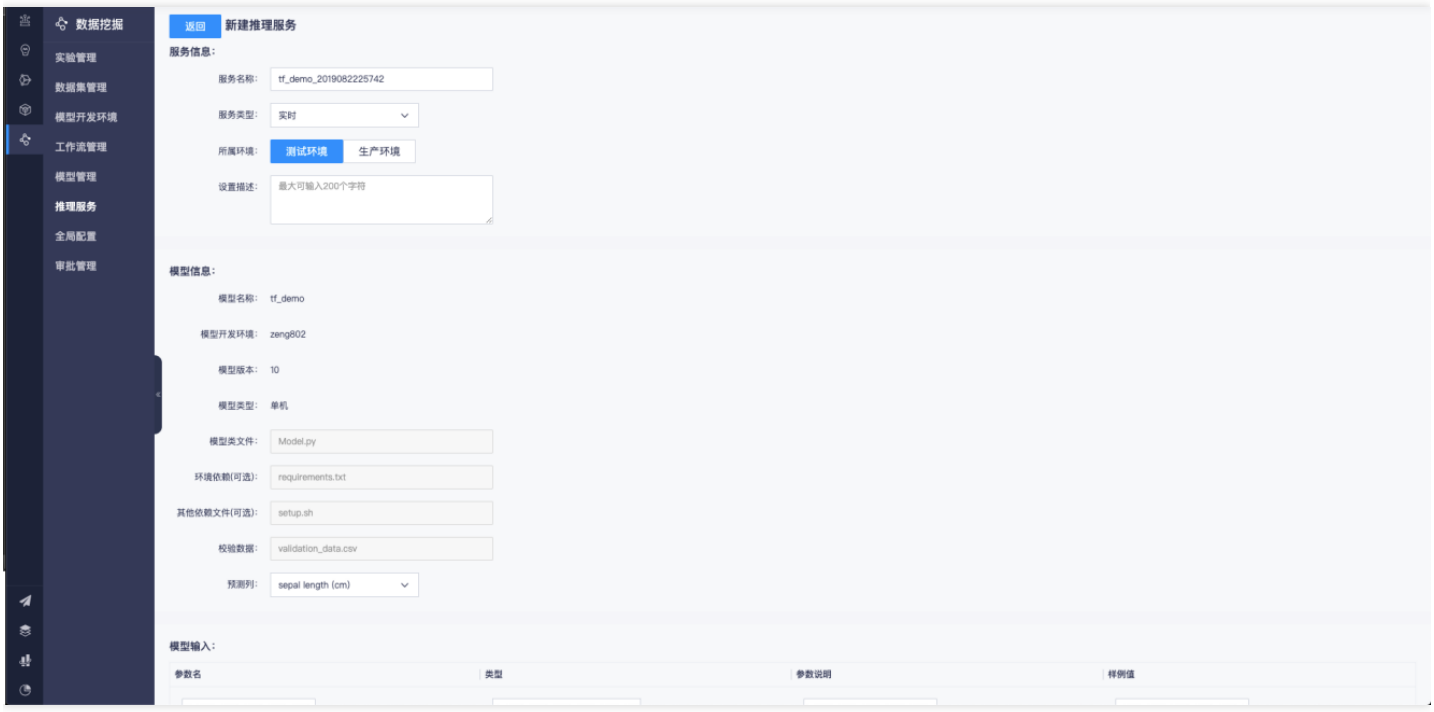


创建好的模型重训练会展示在模型管理>模型重训练的列表中，用户可以对运行中的重训练进行暂停和删除。

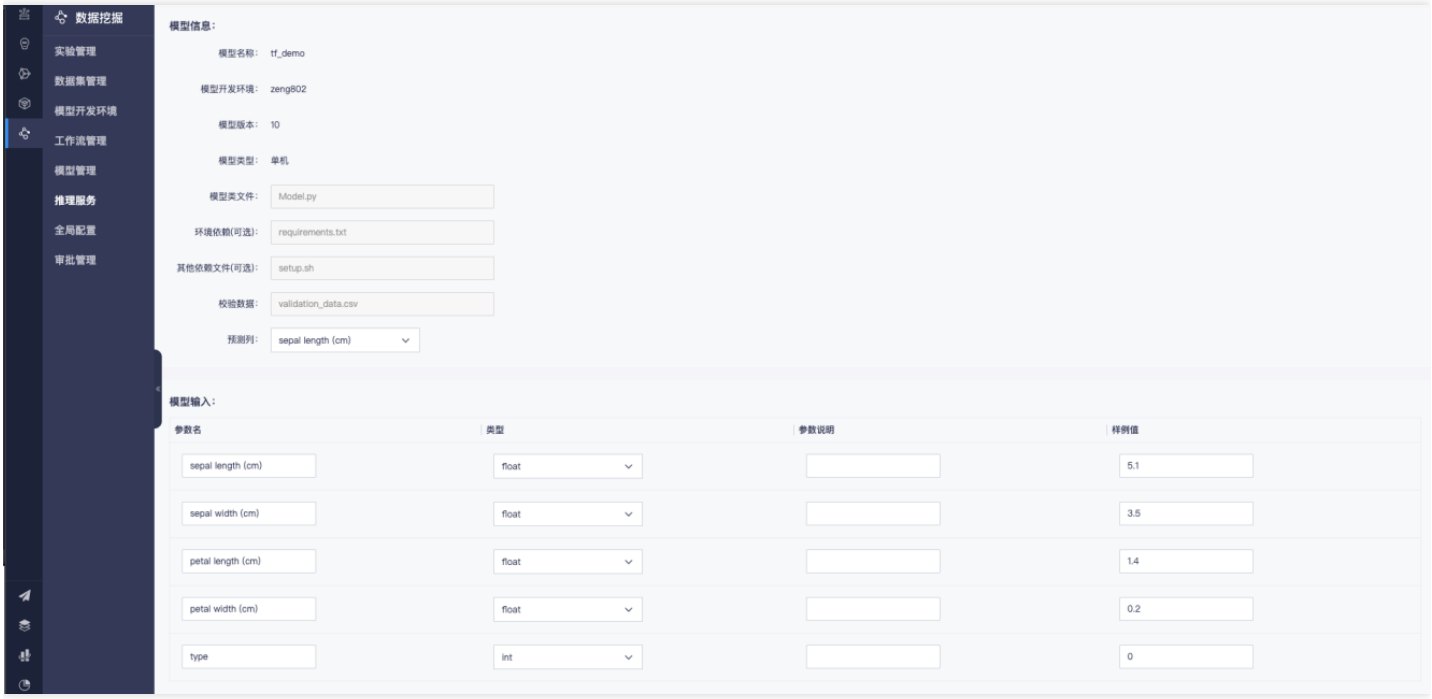
推理服务

最近更新时间: 2019-11-13 06:14:43

在模型列表中创建推理服务，可以把指定模型发布为推理服务。

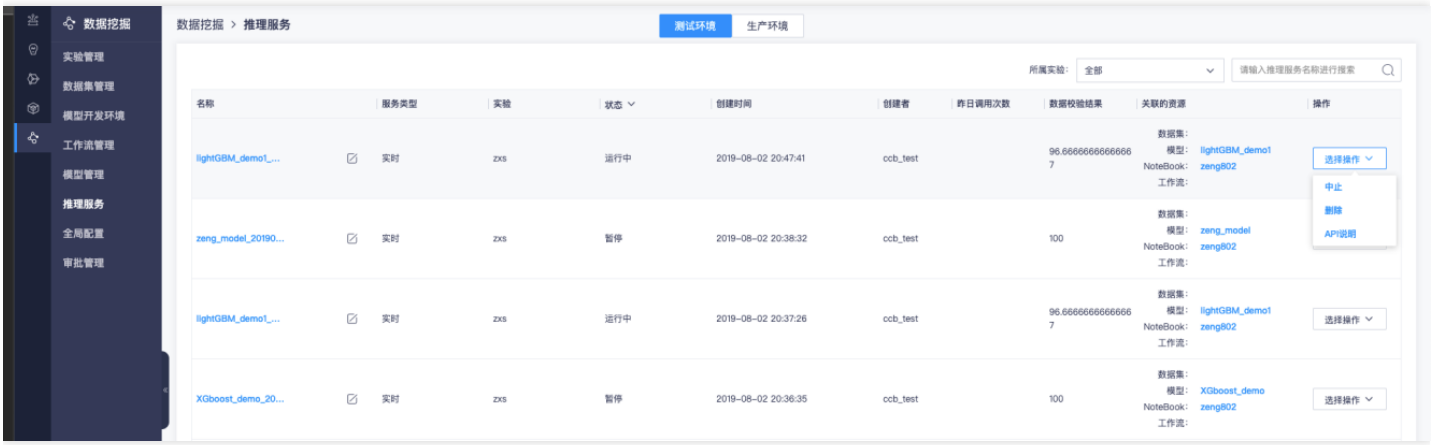


新建推理服务时需要填写模型的相关信息，例如服务的名称、服务类型是实时推理服务还是离线推理服务。



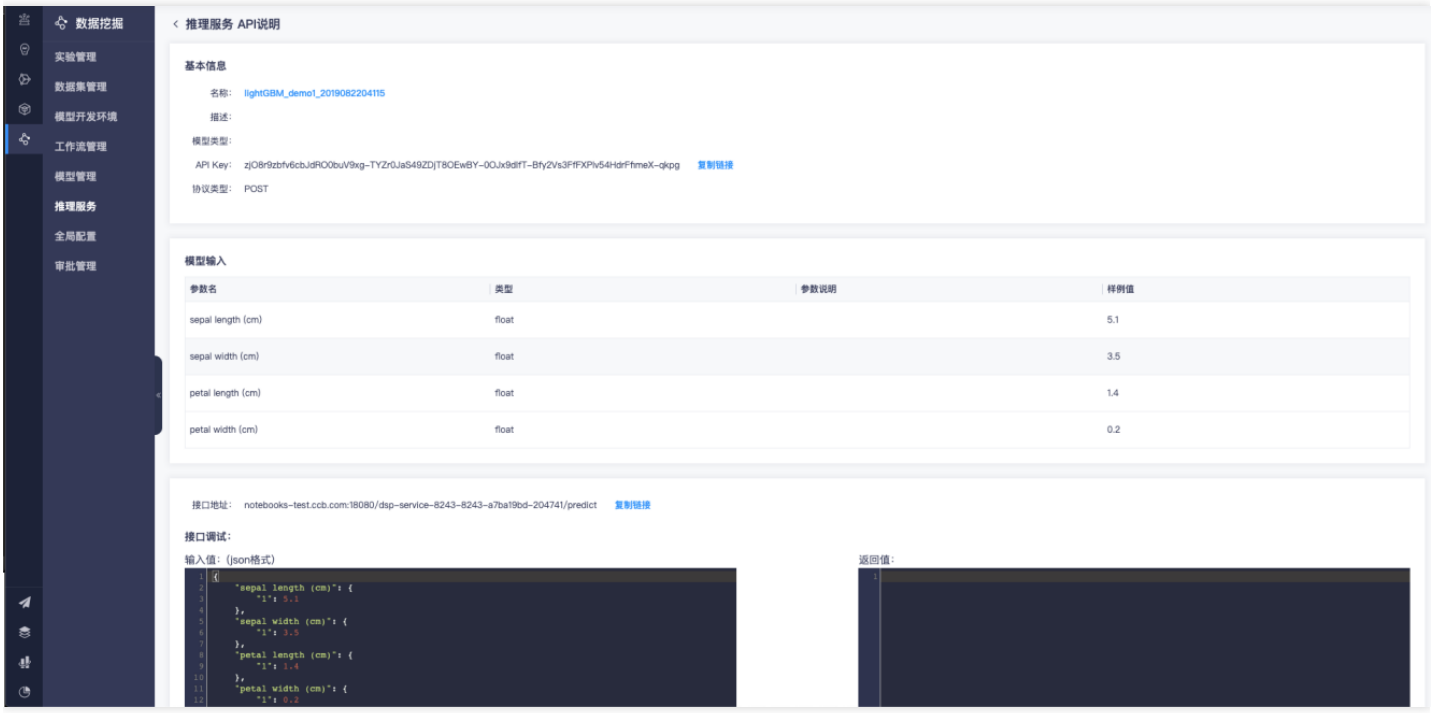
平台通过在模型信息中的校验数据来获得模型输入的schema，预测列以外的字段都会被作为模型输入，可以对输入的类型和样例值参数说明进行修改。

选择推理服务对应的资源信息后会创建对应的推理服务。测试环境中发布推理服务不需要审批，生产环境中发布推理服务需要项目管理员审批。

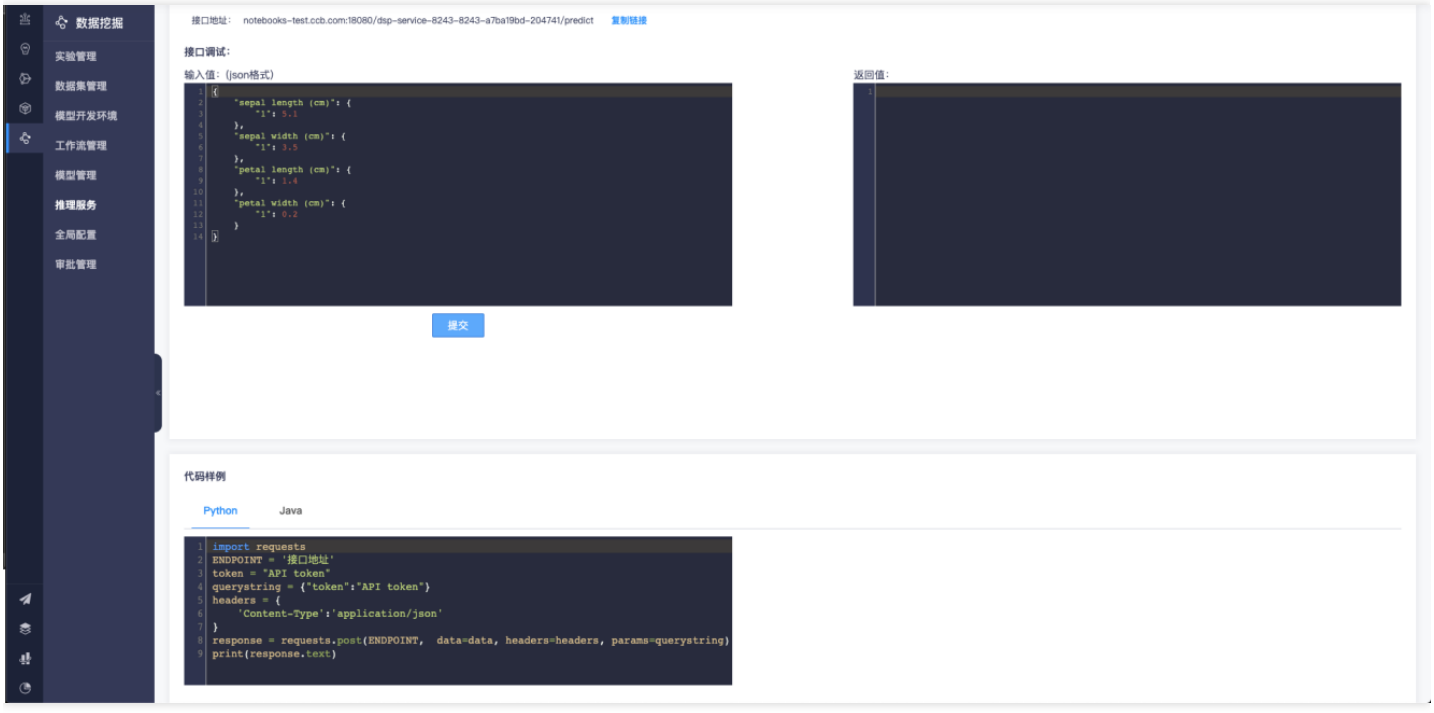


推理服务列表页可以对推理服务进行相关操作： 中止：中止当前推理服务 删除：删除当前推理服务 API说明：查看

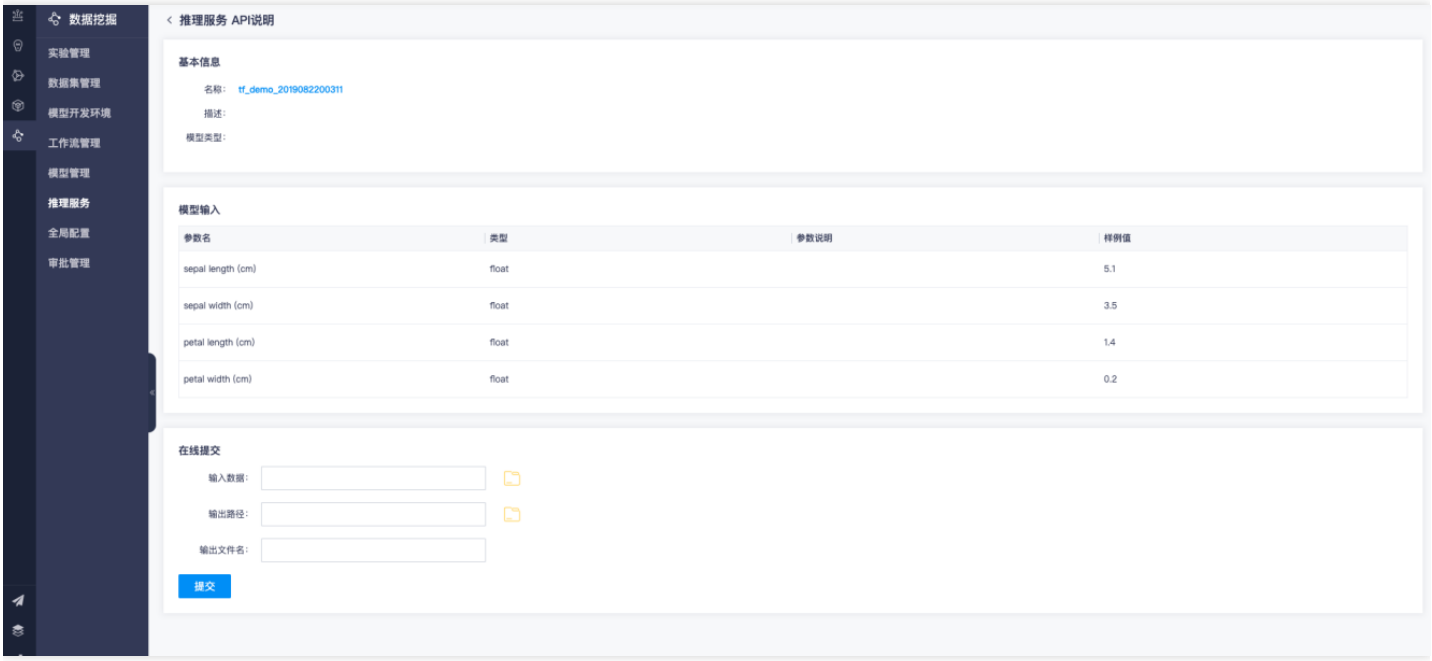
当前推理服务的API说明 推理服务API说明



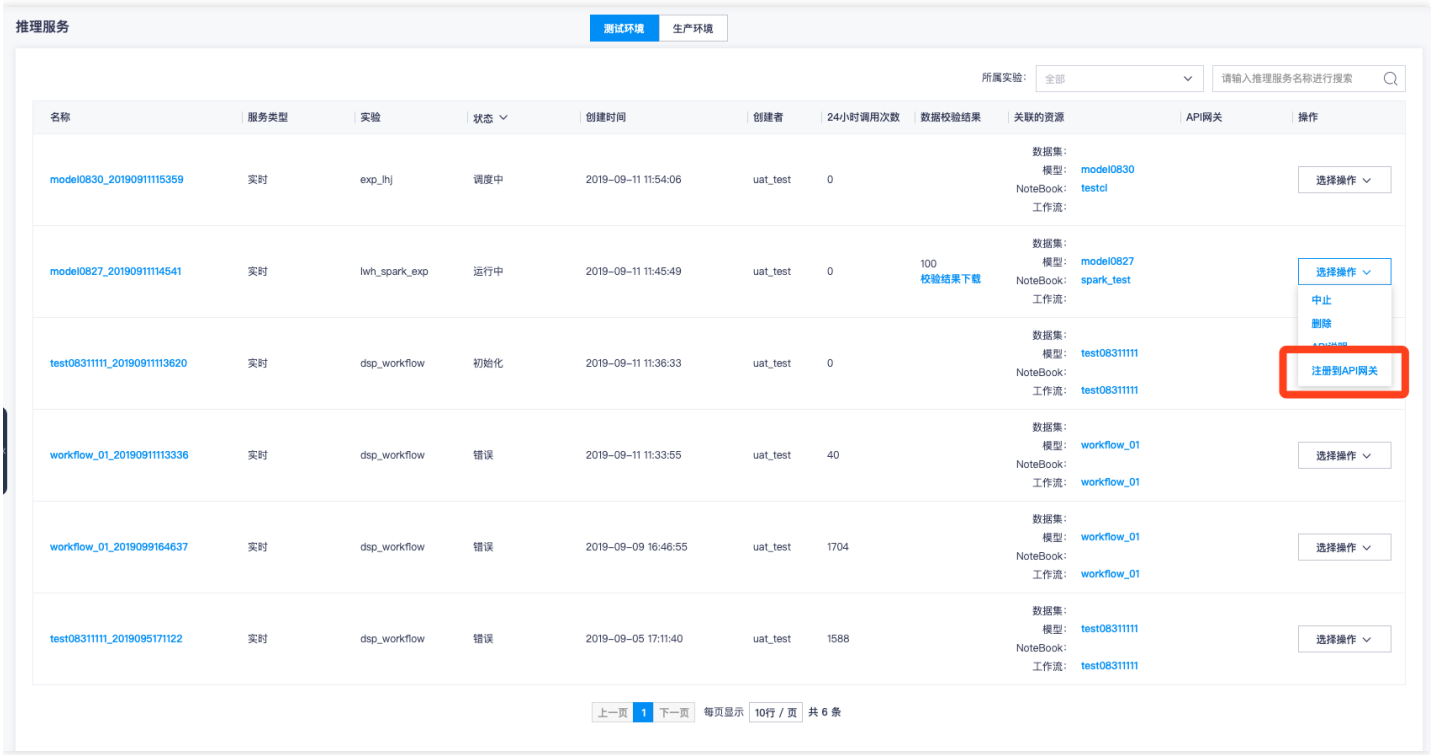
实时推理服务的API说明中包含推理服务的模型输入信息、接口地址信息。



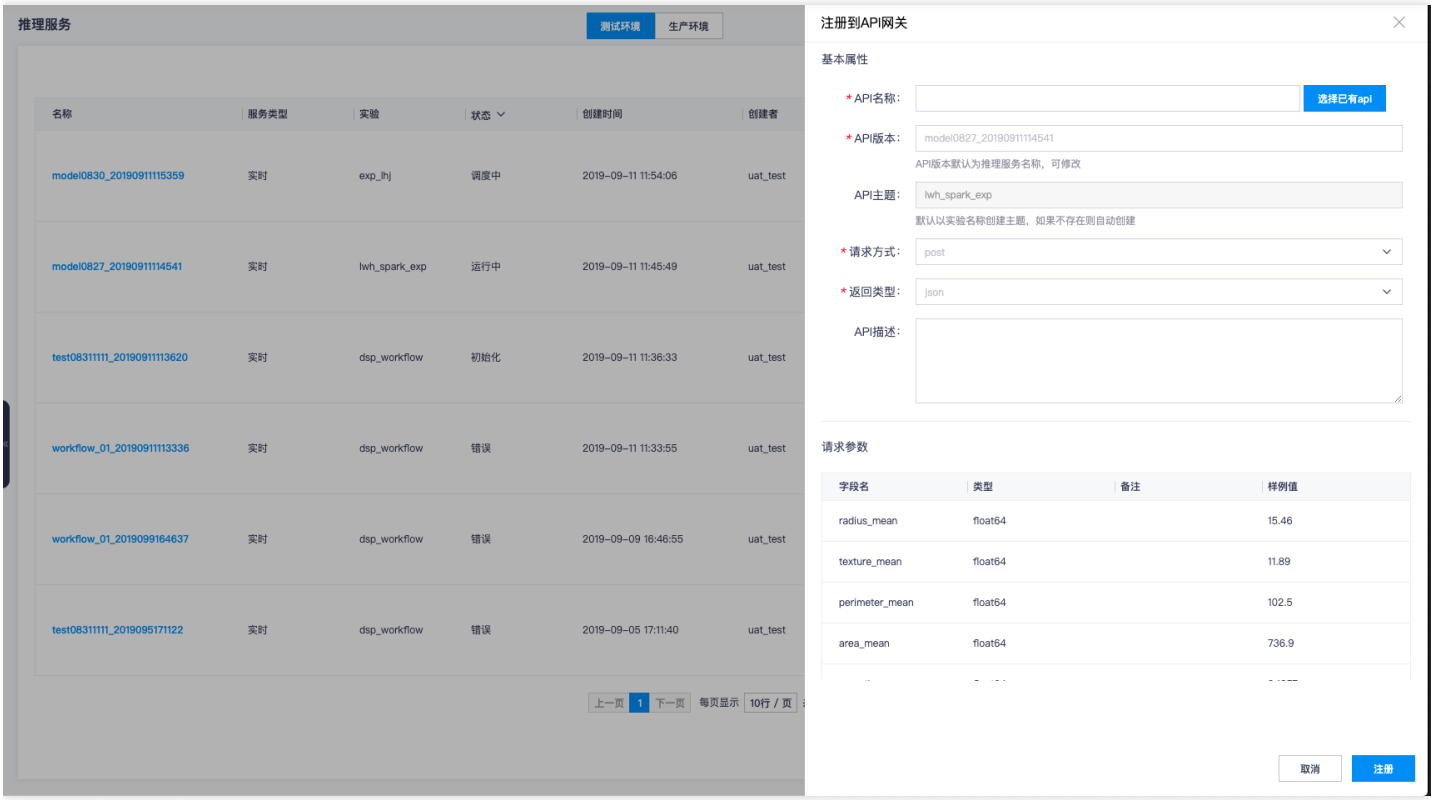
同时也提供了API调试的窗口以及API调用代码样例，用户可以在API说明页面了解实时API的调用方式。



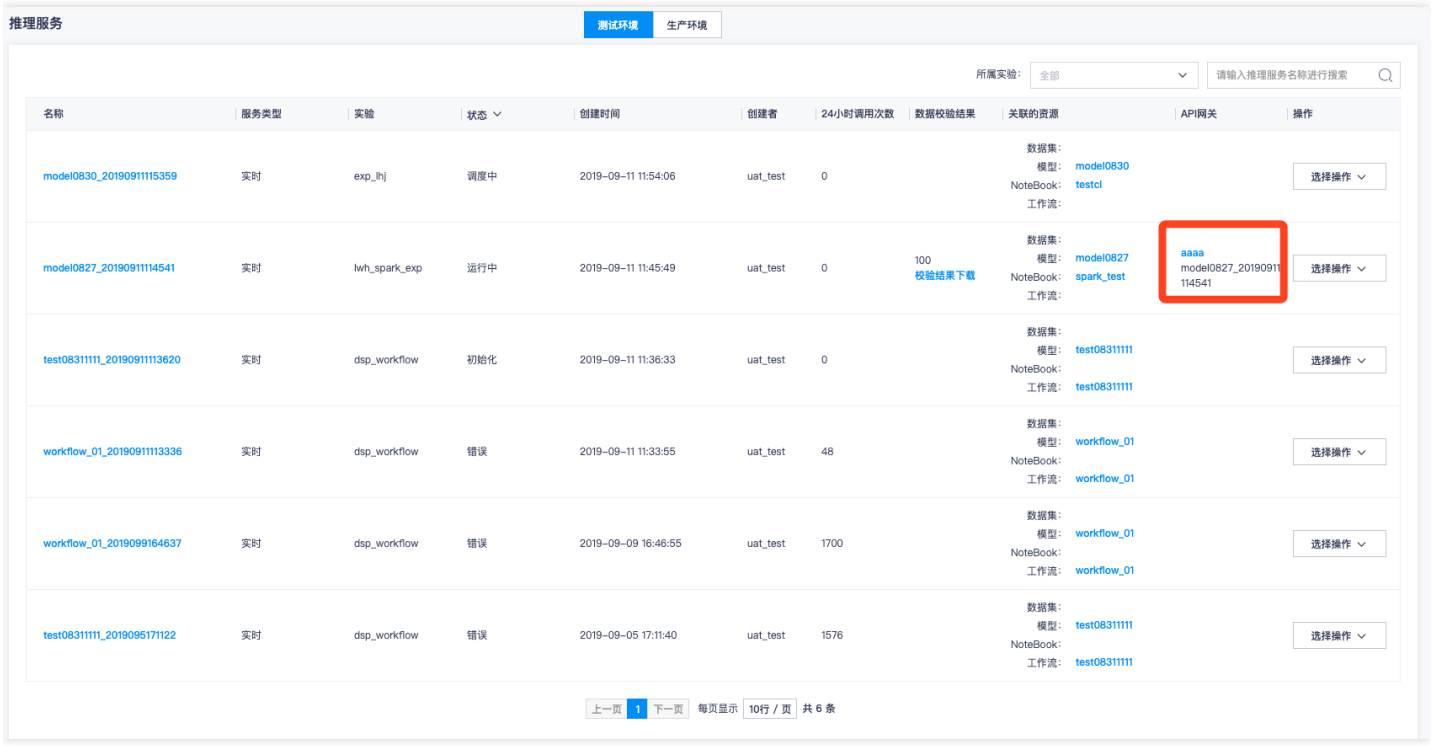
批量推理服务的API说明页面包含在线提交的页面，用户可以在页面上直接使用我的数据中的数据进行离线推理，相关推理结果会写入指定的COS数据目录中。 实时推理服务注册到API网关 数据挖掘组件提供实时推理服务供外部应用程序调用，可以将实时推理服务注册到API网关，相关应用的鉴权需要通过API网关来实现。



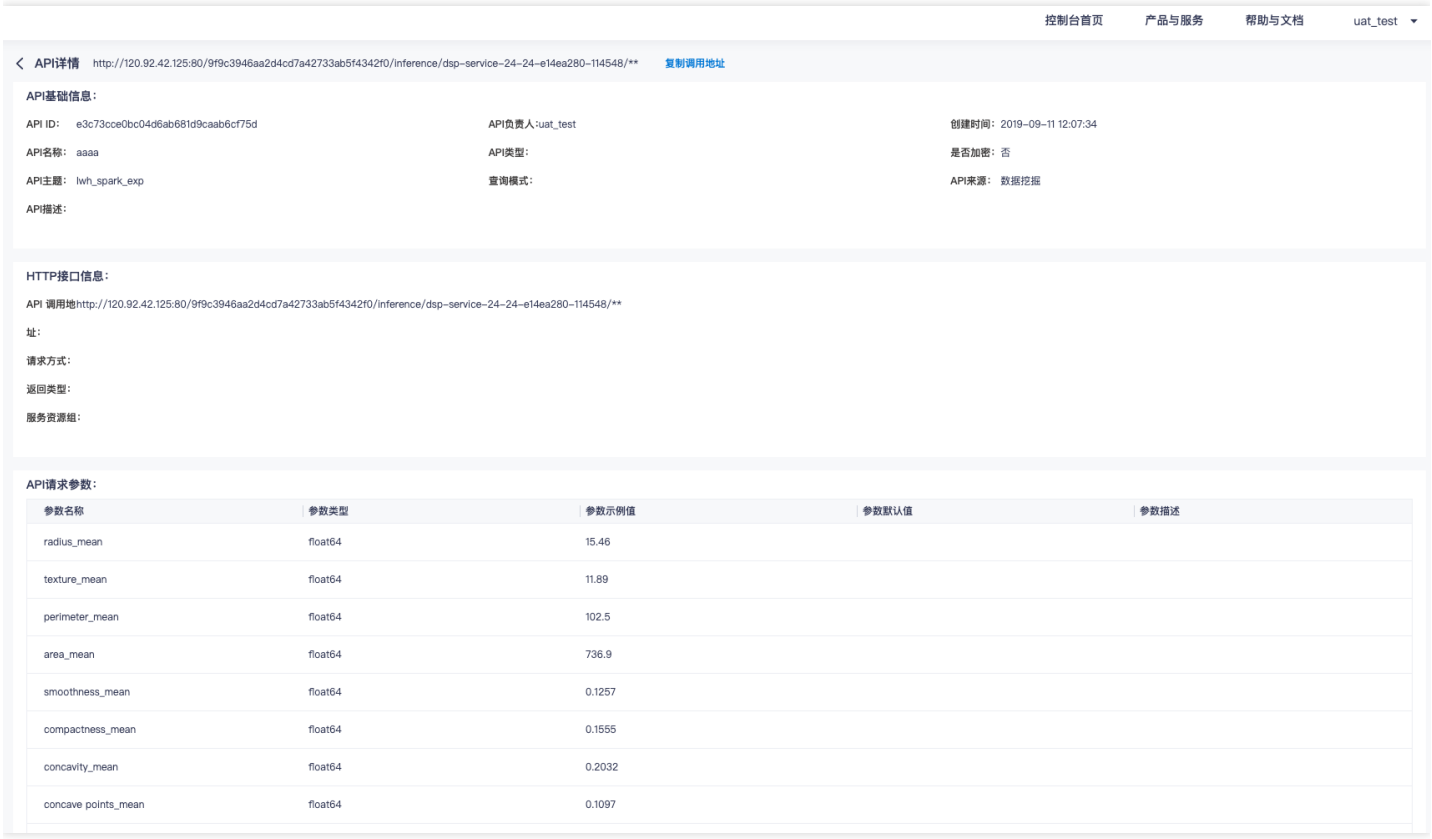
在推理服务列表中，点击实时推理类目右侧的操作，从操作中选择注册到API网关。



填写API网关相关信息。API名称用于标识具体的API，可以选择现有的API也可以新建API。API版本默认为推理服务名称，API主题为实验名称。



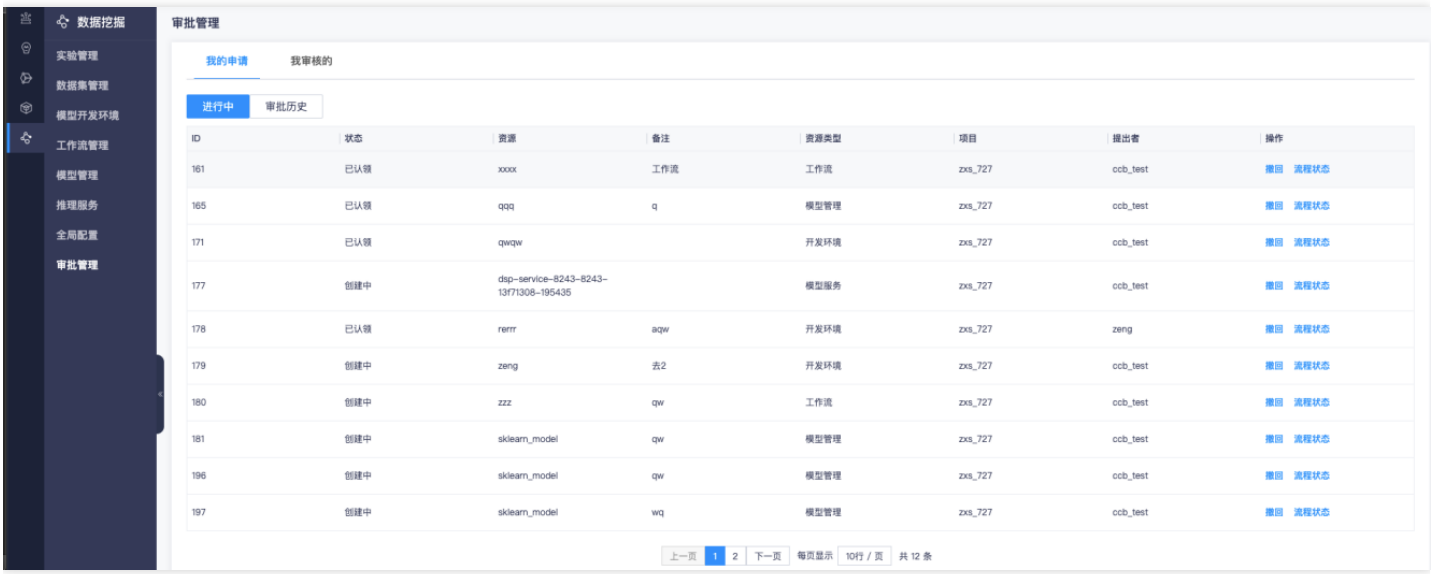
注册成功的API网关会显示在API网关列中，点击API网关名称可以跳转到API网关详情页。



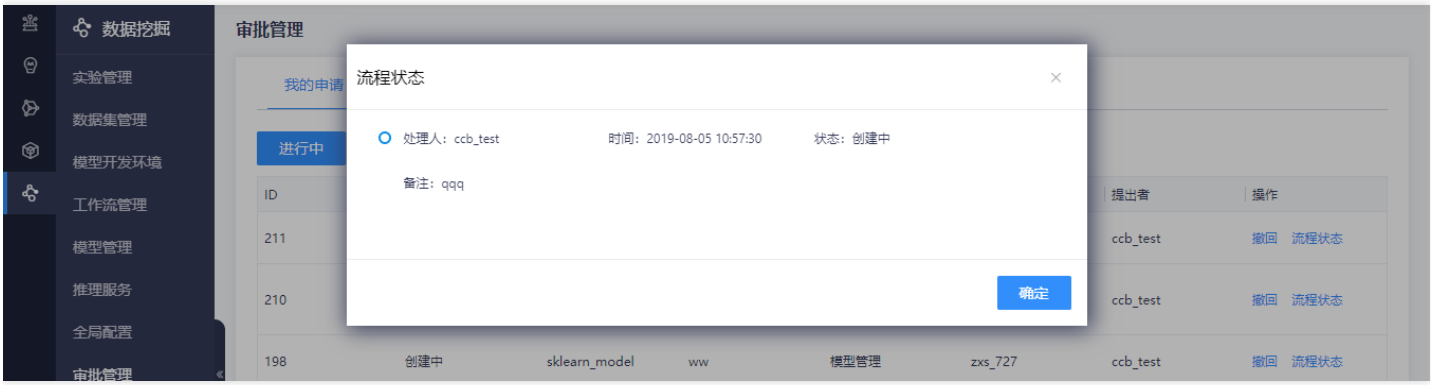
审批管理

最近更新时间: 2019-11-13 06:14:43

我的申请

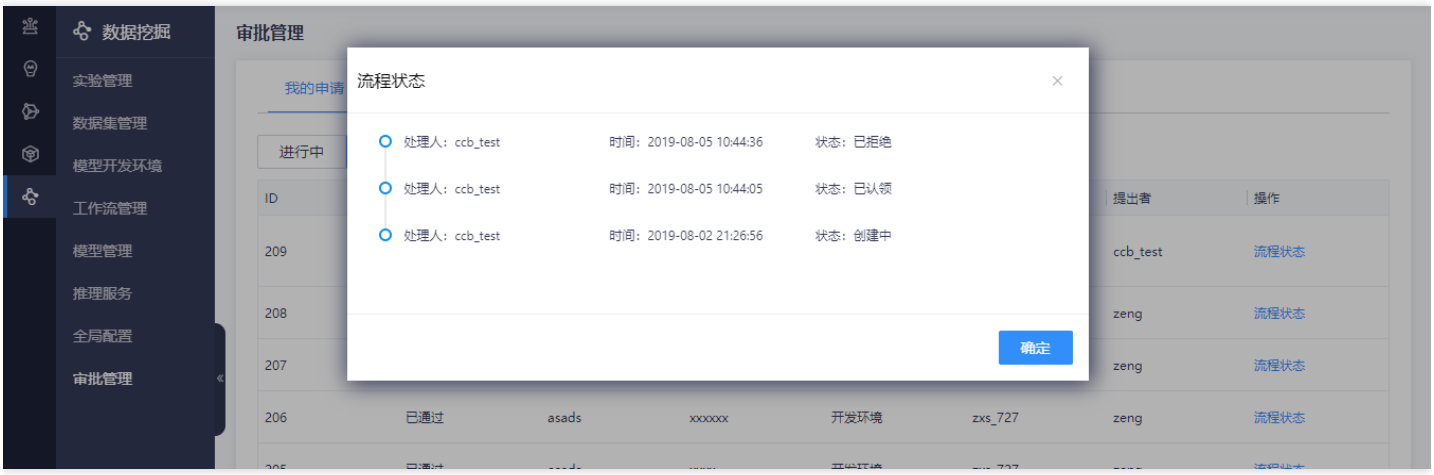


我的申请显示当前用户提交的相关审批申请。用户可以撤回审批申请或查看流程状态。点击查看流程状态，页面显示处理人和时间等信息。

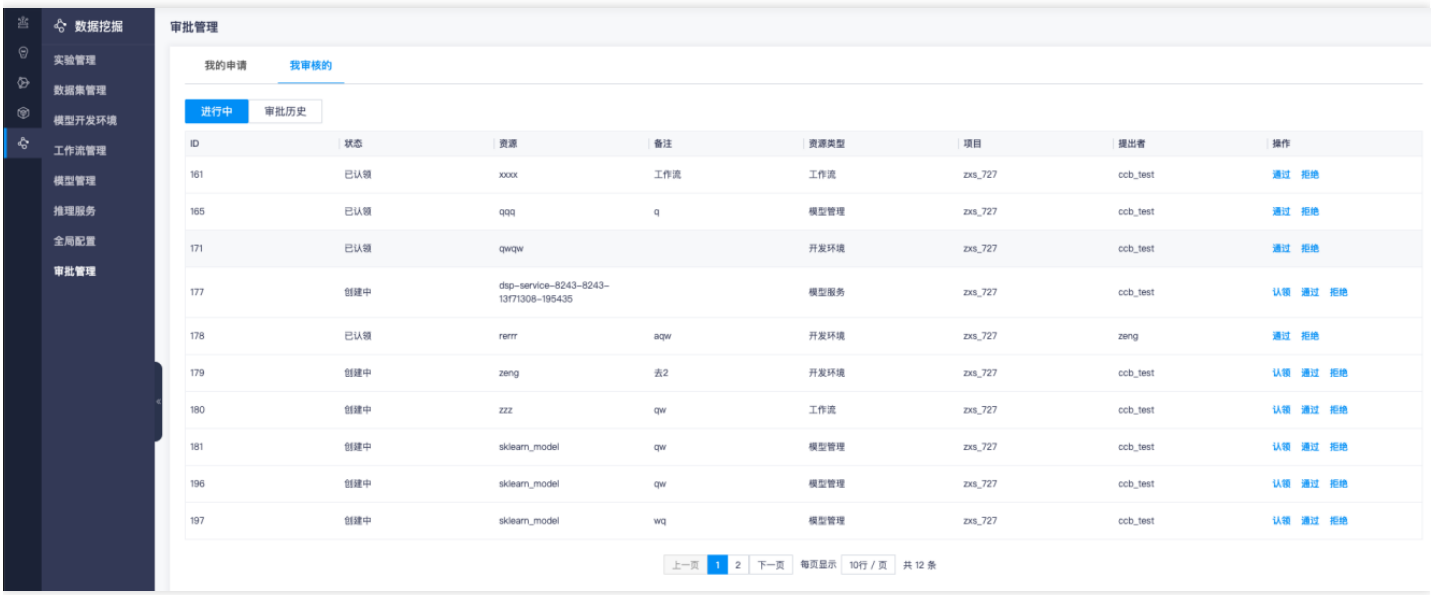


审批历史中显示历史的审批记录。

在审批历史中可以显示流程状态，具体内容为处理人、时间等信息。



我审核的



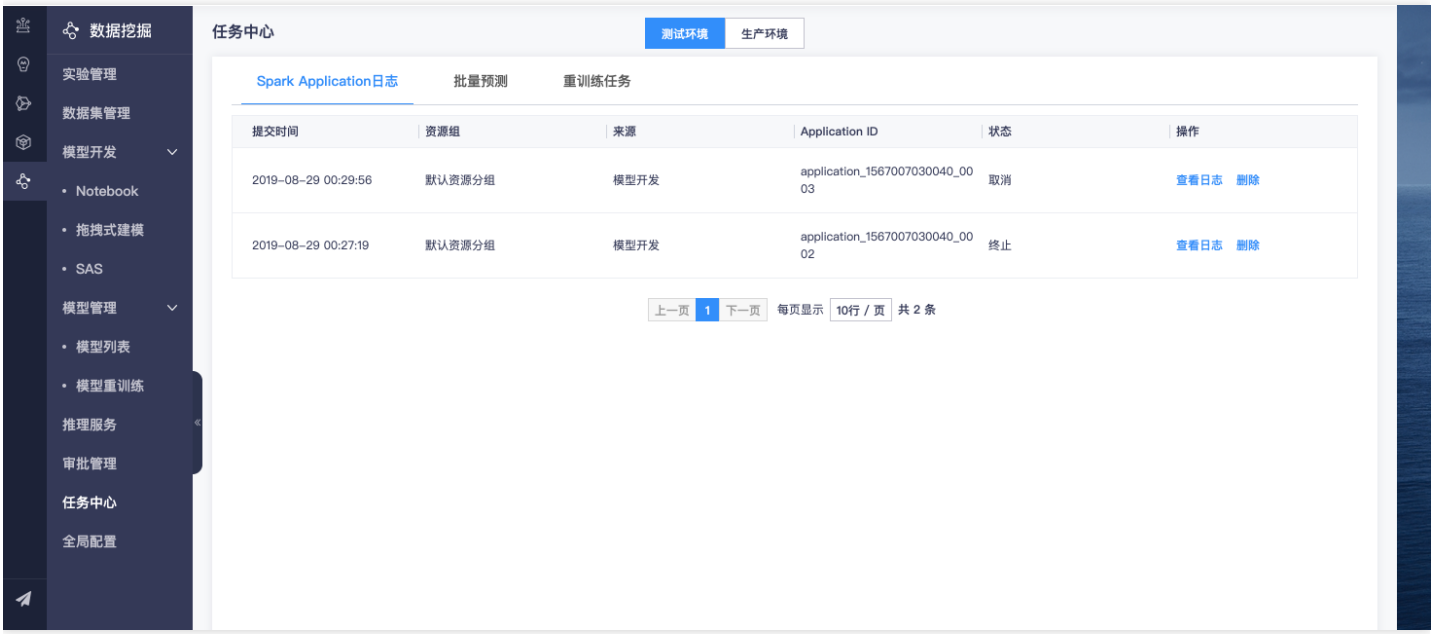
项目管理员会看到当前项目下需要审核的请求。项目管理员可以对审核申请进行认领，认领后当前审批仅当前管理员可见，项目管理员可以对申请进行通过或拒绝。审批历史中显示历史的审批记录。

任务中心

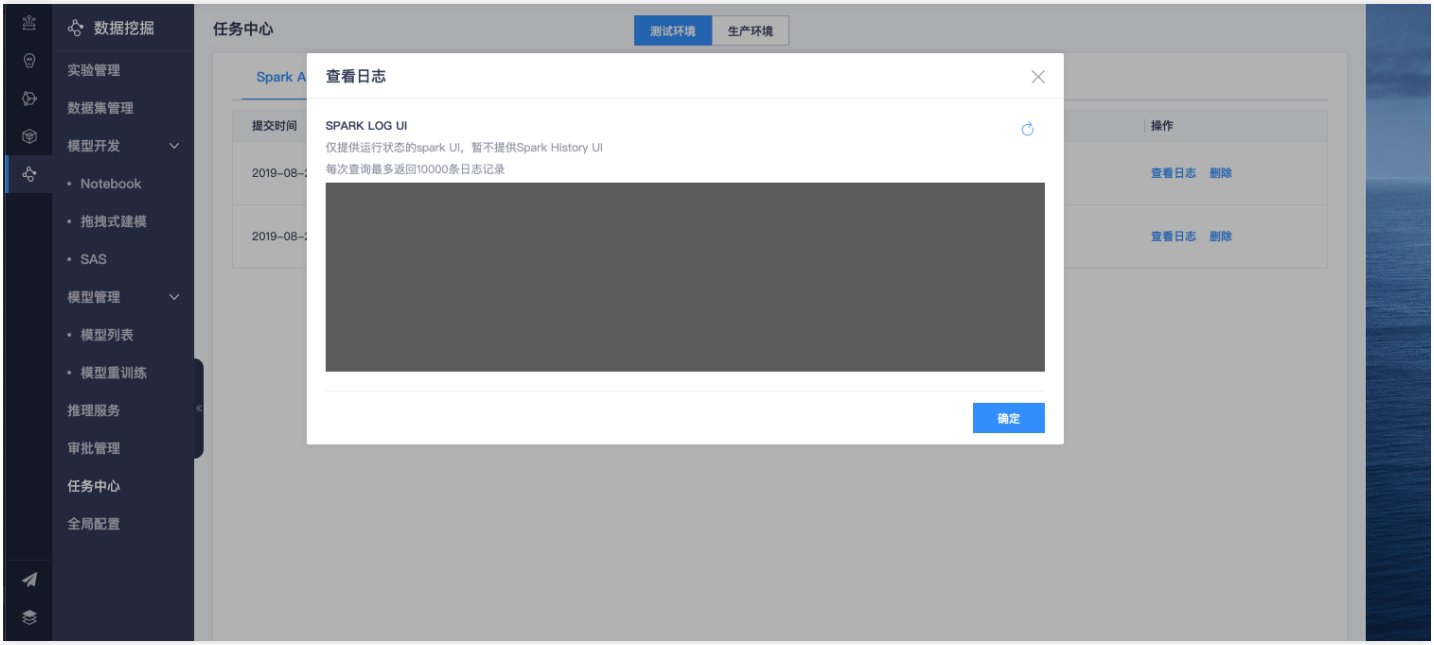
最近更新时间: 2019-11-13 06:14:43

任务中心展示平台中相关的任务信息，包含spark的application，批量预测的任务以及重训练的任务。

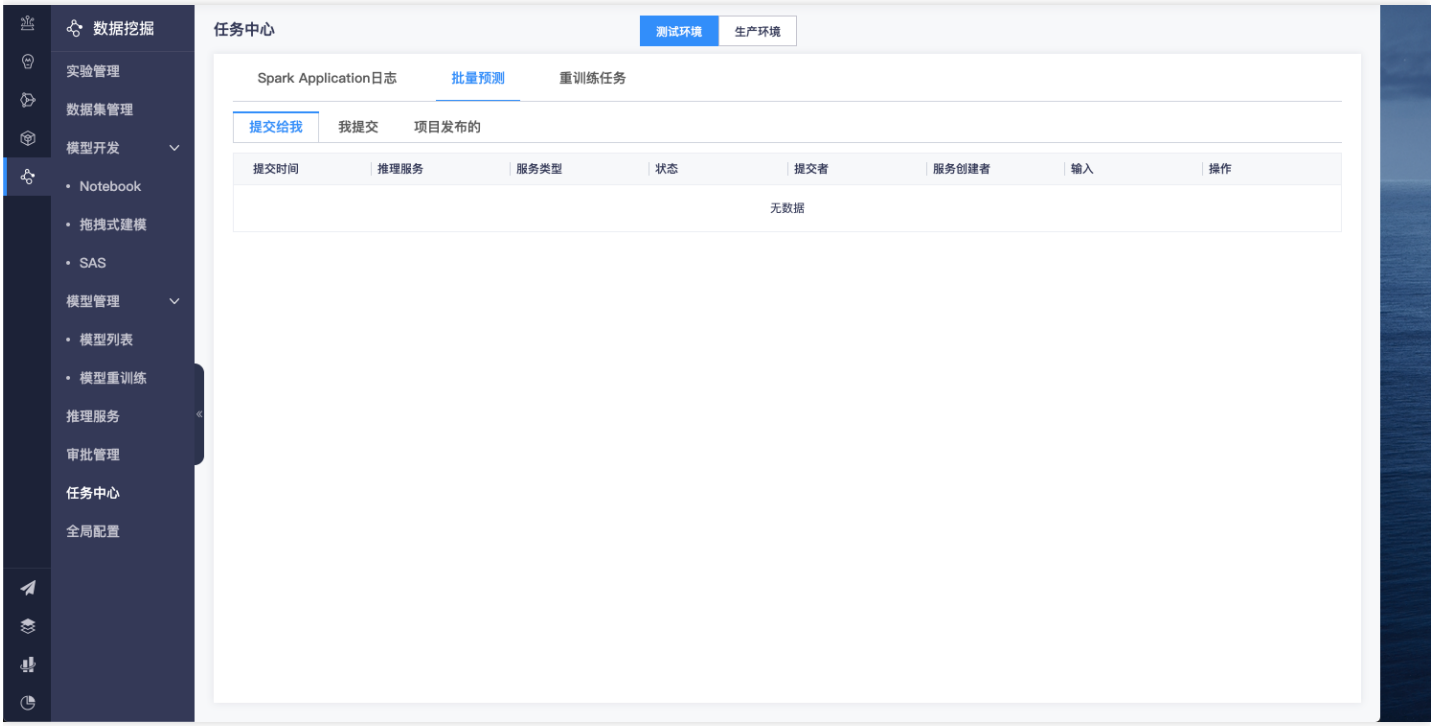
SparkApplication日志



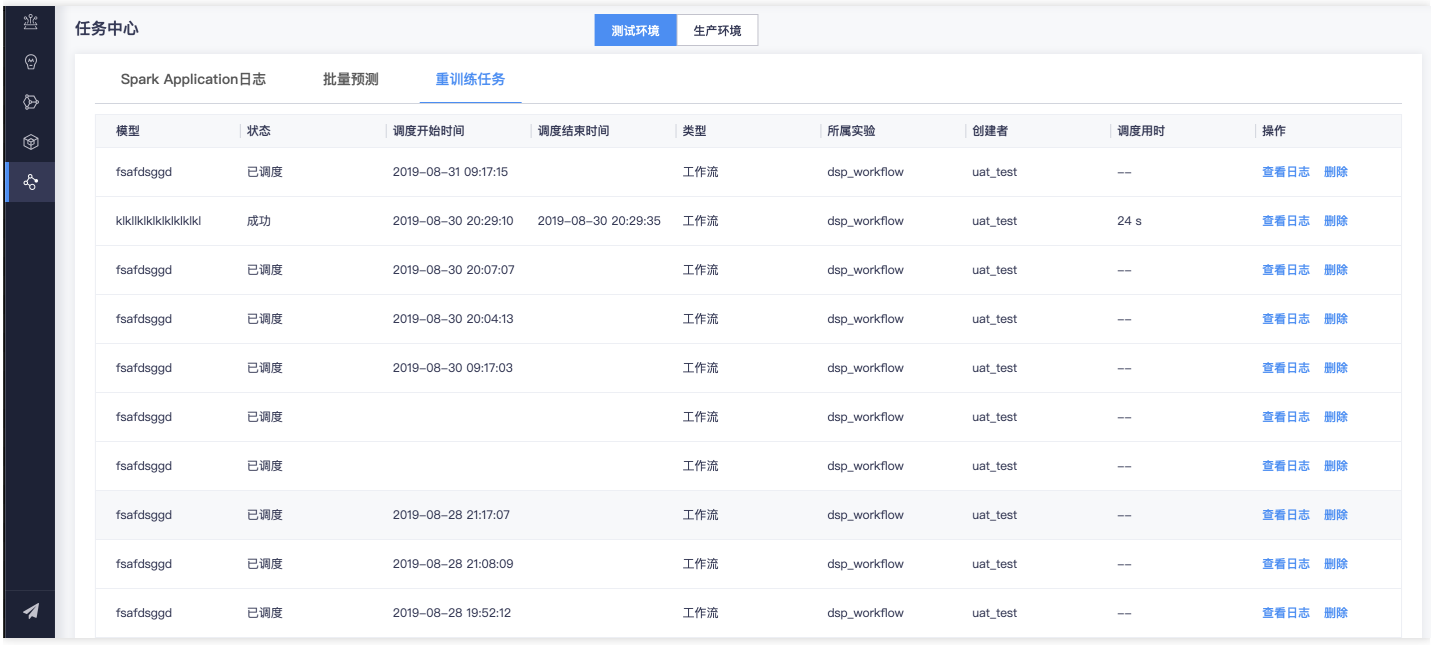
SparkApplication日志展示平台中调用Spark产生的Application相关信息及日志。



点击查看日志会显示当前Spark Application的相关日志。 批量预测



批量预测中展示批量预测相关任务，包含三个页面，项目发布的展示当前项目下发布的所有批量预测任务列表，方便用户查看当前项目下的批量预测服务。 提交给我是指其他用户使用当前用户发布的批量推理服务提交的批量推理任务。 我提交是指我提交的批量预测任务。 重训练任务

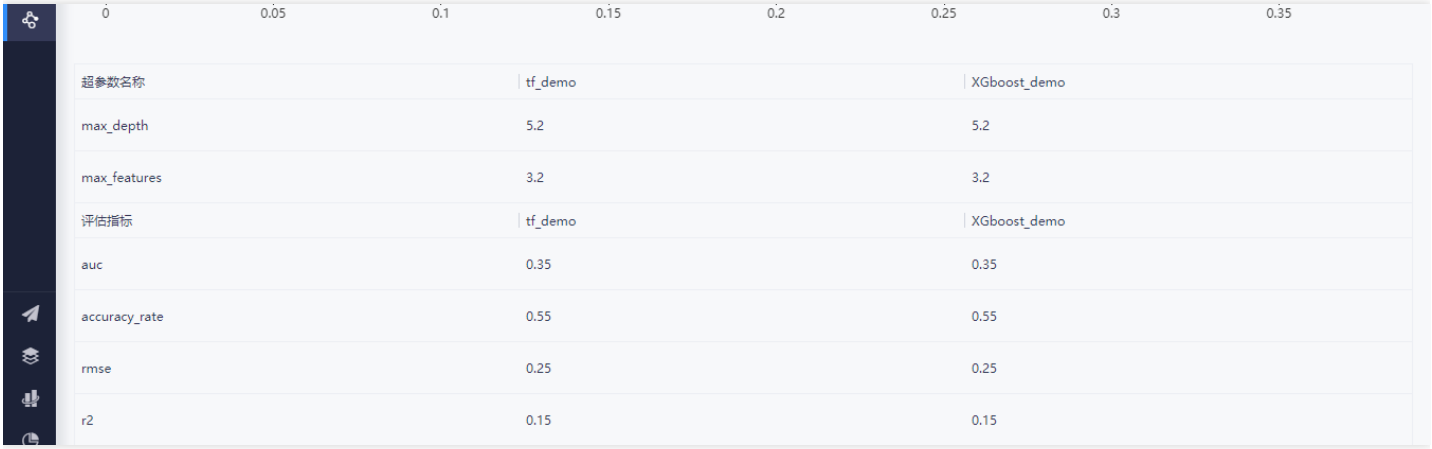


重训练任务页面展示在模型重训练中设置的重训练调度规则，规则实际调度的任务，用户可以查看具体的调度任务或者删除对应的调度任务历史。

全局配置

最近更新时间: 2019-11-13 06:14:43

删除镜像 页面显示用户的自定义镜像，选择镜像可以实现删除操作。



最佳实践

最近更新时间: 2019-11-26 15:30:16

用户身份验证：数据挖掘组件需要使用对象存储资源，因此，需要租户填写了IaaS层的AK/SK和APPID；此外，数据挖掘需要利用用户的AK/SK来验证用户身份，因此，用户需要有用户级的AK/SK才能使用数据挖掘组件的相关功能。

资源组：数据挖掘组件的相关服务需要使用CU和DCU这些计算资源，因此，在开通数据挖掘相关的项目时需要确保为数据挖掘组件创建了对应的资源组，并分配了足够的资源配额。

数据权限：数据挖掘组件可以从数据管理获取元数据，并通过数据服务或SparkMagic读取相关的数据，用户想要使用相关的数据表，需要先在数据管理申请相应的权限。

实时推理服务：数据挖掘组件提供两种推理服务：批量推理和实时推理，批量推理用于解决对时间不敏感（通常需要消耗为几分钟甚至几小时），一次接受的样本量（通常是几万甚至几百万）较大的情况，产出的结果是一个结果名单；实时推理服务用于应对对时间延迟很敏感（通常要求亚秒级的响应），一次推理一个样本。在实时推理场景下，需要依赖API网关对应用调用进行鉴权。

CU/DCU资源申请：数据挖掘组件需要使用CU/DCU这些计算资源来进行模型训练以及对外提供推理服务。对于一个5人左右小团队来说，20CU/DCU左右可以应对一般的工作负荷，如果工作负荷较多可以适当增加CU/DCU配额。DCU是数据挖掘组件的基础计算资源，是K8S相关计算资源的配额，而CU是yarn相关计算资源的配额，只有当需要使用Spark框架运行相关计算任务时才需要申请。

故障指南

最近更新时间: 2019-11-26 15:30:11

Q：新建子账户无法进入挖掘组件。 A：挖掘组件需要依赖对象存储和用户的AK/SK，检查一下租户的IaaS层AK/SK,APPID是否正确配置，此AK/SK是否处于生效状态。此外，数据挖掘组件只允许项目管理员和建模人员两个角色使用相关功能，请确认当前用户的角色是否正确。

Q：无法看到资源组？ A：挖掘组件的需要用到相关资源，创建项目后，项目管理员需要分配对应的资源组和资源。

Q：资源组的剩余资源不足。 A：租户或者项目管理员可以在项目管理中为对应的资源组分配更多的资源，供数据挖掘组件使用。

常见问题

产品介绍常见问题

最近更新时间: 2019-11-26 15:30:11

Q：什么是数据挖掘服务?

A：数据挖掘组件是一个通用的数据挖掘平台，与大数据云其他组件紧密结合，支持主流的机器学习算法，提供单机与集群的机器学习框架支持，结合Spark集群提供分布式内存计算的强大性能，为用户提供模型开发、模型训练、模型部署等一站式数据挖掘服务。

Q：CU和DCU有什么区别?

A：CU对应的是Yarn CU，当使用Yarn计算资源时需要用到。挖掘组件在用到Spark框架时需要占用CU，具体来说，在notebookSpark相关Kernel中创建sparksession时需要指定资源使用；创建集群模型的推理服务时需要指定CU；集群模型的重训练也需要指定CU。而DCU对应的是容器计算资源，在创建Notebook，工作流，推理服务等实例时均需要指定DCU。

产品性能常见问题

最近更新时间: 2019-11-26 15:30:11

Q：我该使用多少CU/DCU? A：根据运行任务的不同需要的CU量不同，用户需要根据自身的实际情况选择对应的CU。 使用Notebook时，运行单机任务时，需要的DCU应该随着任务的复杂情况及要处理的数据量有所增长，而运行Spark任务时，Notebook只是作为发送任务的IDE，不需要太多的DCU。

使用实时推理服务时，可以通过弹性伸缩的方式由平台自动进行资源申请和释放，平台可以感知底层实例的负载，按照既定规则进行资源的水平扩缩容，以应对请求量的突增或突减。

使用批量推理服务时，平台会根据提供的数据量进行自动的数据拆分和合并，不同的CU和DCU会影响任务完成的速度。

CU和DCU之间没有确定的比例关系。对单机框架使用较多的项目中，需要申请更多的DCU，可以不申请CU，而在Spark集群框架使用较多的项目中，需要申请更多CU，申请少量的DCU用于提交任务。

Q：每个Notebook有多少存储空间？磁盘读写性能如何?

A：Notebook使用对象存储作为工作目录和数据目录，实际的存储空间取决于对象存储产品限制，理论上可以无限横向扩展。

由于对象存储是web文件系统，尽管数据挖掘组件与对象存储是通过高速内网相连，数据需要通过网络传输，因此，实际磁盘读写速度还是会比本地磁盘慢一些，但对数据挖掘任务来说影响甚微。

产品使用常见问题

最近更新时间: 2019-11-26 15:30:11

Q：Notebook预置了哪几种镜像都包含哪些框架？ A：目前Notebook预置了四个镜像。 通用：包含大部分python数据科学的数据科学库同时支持pyspark，也包含了tensorflow等深度学习的python库，包含的主要数据科学库如下：pandas, matplotlib, numpy, seaborn, scipy, statsmodels, daskskearnxgboostlightgbmpyspark r tensorflow(cpu) caffeMXNetpytorch 单机机器学习：包含大部分python单机数据科学的库，包含的主要数据科学库如下：pandas, matplotlib, numpy, seaborn, scipy, statsmodels, dask, sklearn, xgboost, lightgbm, r 集群机器学习：包含集成了sparkmagic的pyspark kernel，可以使用pyspark集群进行分布式数据挖掘，同时也集成了主要的python数据科学库，包含的主要数据科学库如下：pyspark, pandas, matplotlib, numpy, seaborn, scipy, statsmodels, dask, sklearn, xgboost, lightgbm, r 深度学习（cpu）：包含主流的python深度学习的库，包含的主要数据科学库如下：pandas, matplotlib, numpy, seaborn, scipy, statsmodels, dask, tensorflow, caffe, MXNet, pytorch (cpu) 同时平台也支持用户在创建镜像时，通过re**Quirement.txt以及setup.sh这些环境初始化脚本来定制自己的环境，用户定义的开发环境可以通过自定义镜像功能保存为用户自定义镜像，方便下次使用。

Q：为什么资源组显示还有资源而提交的任务却会因没有资源而失败？ A：出现这种情况可能有以下几种可能： 资源组的剩余资源检测有15S的延迟，提交任务时底层已经没有资源了，或项目管理员缩减了资源组的资源。故任务可以提交但因为缺少资源而初始化失败。 另外，剩余资源也可能被批量预测和工作流占用。批量预测和工作流是任务式的资源占用，有相关任务的时候才会占用资源。 出现以上情况可以稍待工作流和批量预测任务完成后释放资源，也可以通过租户或项目管理员账号新增资源组或增加资源组配额。

Q：我发布的推理服务/发布到生产环境的资源为什么看不到？ A：大数据云支持企业级的流程管理，对生产线上资源有影响的操作都需要项目管理员的审批。在生产环境下发布推理服务或将相关任务发布到生产环境都需要经过项目管理员审批，相关的审批进度可以在审批管理查看。

Q：在哪里可以看到我的程序日志？ A：挖掘组件中提供两类日志，一类是K8S的POD日志，一类是Spark的日志。具体来说，用户可以在以下页面看到相关日志信息。 Notebook：用户可以在Notebook中查看具体的日志 工作流：用户可以在左上的运行历史中查看每次运行的日志。 推理服务：在推理服务详情页可以看到推理服务的日志。 此外，在任务中心中，可以按照SparkApplicaition查看每个Applicaiton的日志；批量预测可以查看批量预测每个任务的运行日志；重训练任务可以查看具体重训练任务的日志。

Q：当前项目下没有任何占用资源的项目，为什么资源组的资源可用资源为0？ A：资源组与项目是多对多的关系，当前资源组的资源被其他项目占用，会导致资源组配额减少甚至为0，此时可以稍待资源组资源释放或者向相关权限账号申请新建新的资源组、增加原有资源组的配额。