ALIBABA CLOUD

Alibaba Cloud

机器学习PAI
快速入门

文档版本：20220519

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击**设置**> **网络**> **设置网络类型**。
粗体	表示按键、菜单、页面名称等UI元素。	在**结果确认**页面，单击**确定**。
Courier字体	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

1.入门概述

本文指引您快速地构建一个完整的PAI工作流。

PAI-Designer支持模板工作流、自定义工作流两种方式构建工作流。

以下分别为您介绍使用自定义工作流和模板工作流时，从创建到运行成功的完整案例。

- 自定义工作流demo

  使用自定义工作流方式创建实验时，Designer为您提供百余种算法组件，并支持接入MaxCompute表数据或OSS数据等多种数据源，提高建模效率，demo流程如下。

  i. 新建自定义工作流

     将原始数据上传至MaxCompute或OSS中，并配置工作流数据源。

  ii. 数据准备与预处理

     对原始数据进行预处理，生成模型训练集和模型预测集。

  iii. 数据可视化

     对源数据或中间结果数据进行可视化处理，获取数据分析结果。

  iv. 算法建模

     使用符合业务场景的算法组件加上预处理后的数据训练集进行算法建模。

  v. 评估模型

     使用训练好的模型对预测数据集进行结果预测，并结合预测集中的“正确答案”评估模型效果。

- 模板工作流demo

  通过模板可以直接套用模板快速创建工作流，运行成功后进行模型部署，demo详情请参见模板工作流demo。

2.自定义工作流demo

2.1. 新建自定义工作流

Designer通过工作流的方式来实现建模与模型调试，您需要先创建一个工作流，再根据建模需求在工作流中排布不同组件的处理调度逻辑。文本为您介绍如何新建工作流。

前提条件

已开通PAI并创建了工作空间。

背景信息

Designer为您提供多种工作流模板，您可以根据业务需求直接套用工作流模板，也可以结合自身业务全新新建一个空白工作流。本示例以新建一个空白工作流为例，为您介绍新建工作流的操作步骤。

操作步骤

1. 进入PAI-Designer页面。
   i. 登录PAI控制台。
   ii. 在左侧导航栏单击**工作空间列表**，在工作空间列表页面中单击待操作的工作空间名称，进入对应工作空间内。
   iii. 在工作空间页面的左侧导航栏选择**模型开发和训练 > 可视化建模（PAI-Designer）**，进入Designer页面。
2. 新建工作流。
   i. 单击**新建工作流 > 新建空白画布**。
   ii. 在弹出的新建画布页面中配置工作流参数，完成后单击**确定**，完成工作流新建。

本示例中，您必须填写新建的工作流名称，其他参数保存默认状态即可。

3. 在工作流列表页面，选中刚刚新建的工作流后，单击**进入工作流**即可进入工作流页面。

后续步骤

完成新建工作流后，需要进行数据准备与预处理，详情请参见数据准备与预处理。

2.2. 数据准备与预处理

通常情况下，构建一个模型时，您需要将模型构建调试所需要使用的数据准备好并完成数据的预处理，便于后续根据业务需求进行模型开发所需的进一步加工。本示例使用PAI为您提供的公开数据为例，为您示例数据准备与预处理的操作步骤。

前提条件

已新建一个工作流，详情请参见新建自定义工作流。

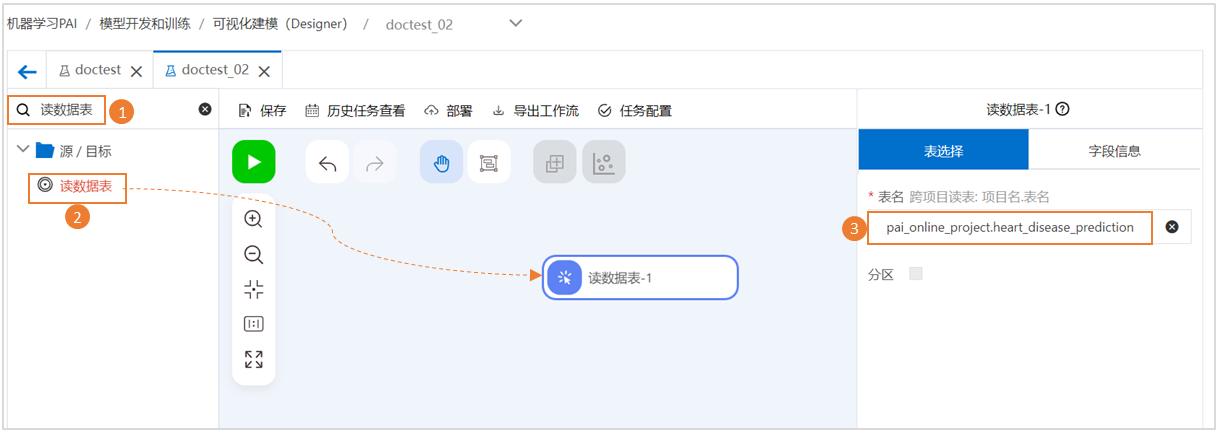
step1：进入工作流页面

1. 登录PAI控制台，在左侧导航栏单击**工作空间列表**，单击待操作的工作空间名称，进入对应工作空间内。
2. 在工作空间页面的左侧导航栏选择**模型开发和训练 > 可视化建模（PAI-Designer）**，进入Designer页面。
3. 在工作流列表页面，选中对应工作流后，单击**进入工作流**即可进入工作流页面。

step2：准备数据

本示例使用PAI为您提供的心脏病案例的公开数据，您无需新建表、写入表数据，可直使用接**读数据表**来读取此公开数据，作为数据准备。

> **说明** 通常情况下，您需要准备一个MaxCompute表或OSS表，并通过**源/目标**组件下的**读数据表**、**写数据表**、**读OSS数据表**等组件来查询或写入数据到表中，操作详情可参见源/目标对应组件文档。



1. 在左侧组件列表的搜索框中，搜索**读数据表**，找到**读数据表**组件。
2. 将**读数据表**组件拖入右侧画布中，画布中自动生成一个名称为**读数据表-1**的工作流节点。
3. 在画布中选中**读数据表-1**节点，在右侧节点配置页面中的**表名**中输入`pai_online_project.heart_disease_prediction`，读取公开的心脏病案例的数据表。

   您也可以将右侧的节点配置页面切换到字段信息页签，查看此公开数据的字段详情。

step3：数据预处理

本示例使用的原始数据为公开的心脏病案例的数据，数据预处理以将所有字段取值归一化为例，为您示例数据预处理的步骤。本示例的数据预处理主要包括三个处理流程：

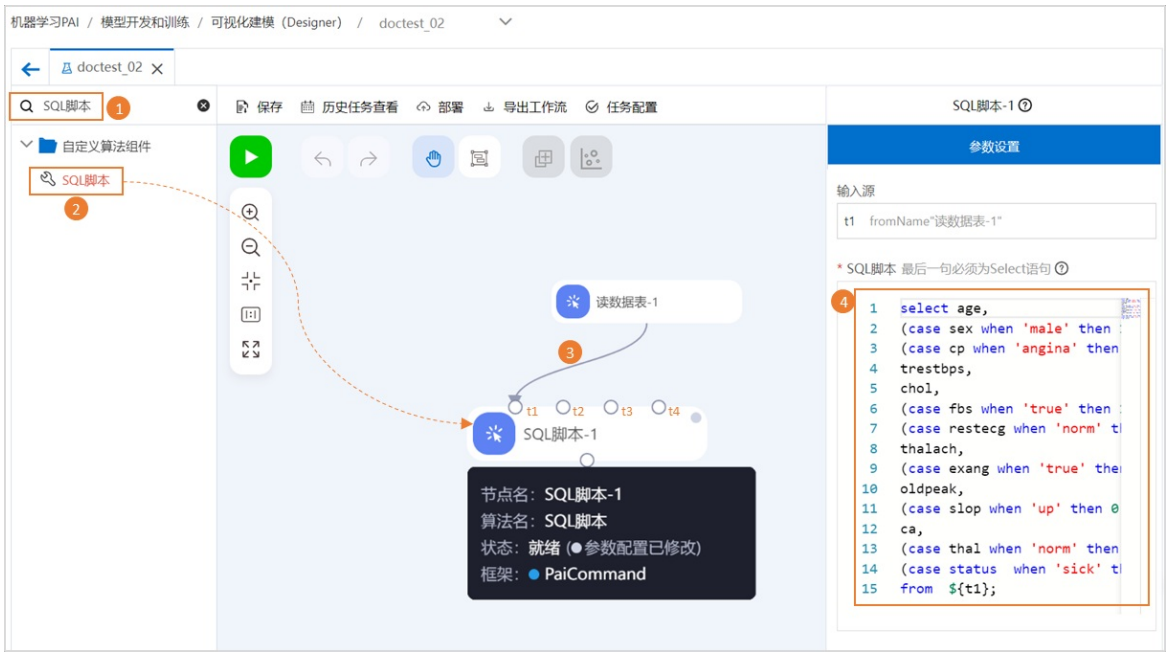
1. 将原始数据表中，取值为非数值类型的字段，通过SQL替换为数值类型的取值，以保障此预处理后，表中所有字段的取值都是数值类型的取值。
2. 将表中所有字段均换为double类型，以保障此预处理后，表中所有字段均满足后续归一化处理的字段属

性要求。

3. 将表中所有字段的取值均归一化。

详细操作如下。

1. 数据预处理：转换非数值类型字段。



i. 在左侧组件列表的搜索框中，搜索**SQL脚本**，找到**SQL脚本**组件。

ii. 将**SQL脚本**组件拖入右侧画布中，画布中自动生成一个名称为**SQL脚本-1**的工作流节点。

iii. 通过连线，将**SQL脚本-1**节点作为**读数据表-1**节点的下游节点。

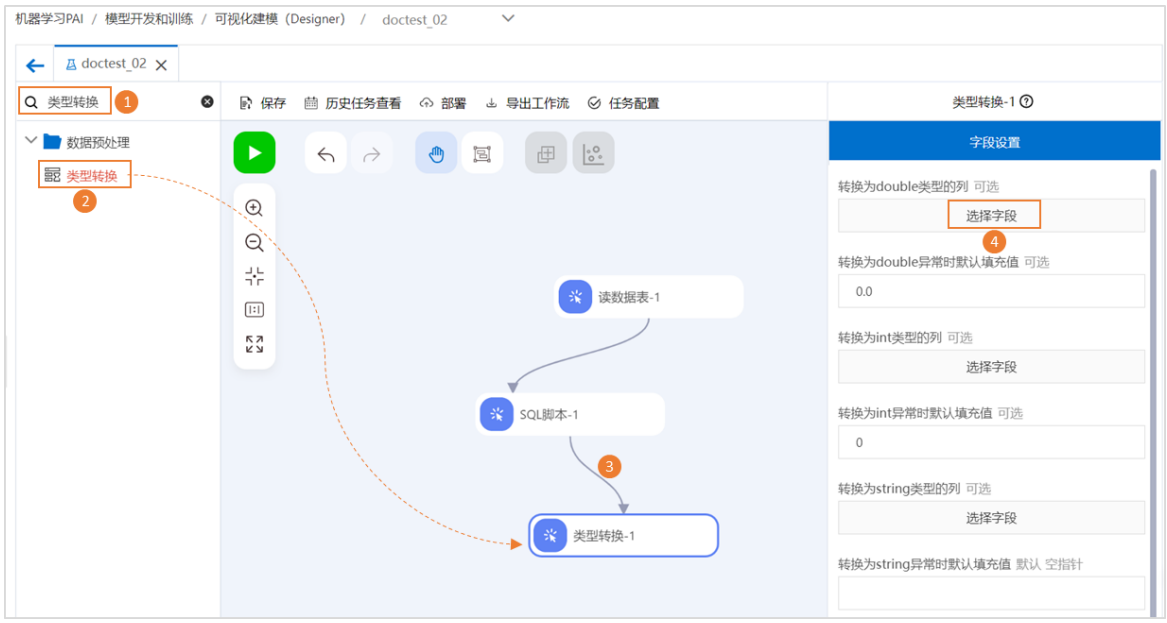
iv. 在画布中选中**SQL脚本-1**节点，在右侧节点配置页面中的SQL脚本输入框中输入下面的SQL代码。

```
select age,
(case sex when 'male' then 1 else 0 end) as sex,
(case cp when 'angina' then 0  when 'notang' then 1 else 2 end) as cp,
trestbps,
chol,
(case fbs when 'true' then 1 else 0 end) as fbs,
(case restecg when 'norm' then 0  when 'abn' then 1 else 2 end) as restecg,
thalach,
(case exang when 'true' then 1 else 0 end) as exang,
oldpeak,
(case slop when 'up' then 0  when 'flat' then 1 else 2 end) as slop,
ca,
(case thal when 'norm' then 0  when 'fix' then 1 else 2 end) as thal,
(case status  when 'sick' then 1 else 0 end) as status
from  ${t1};
```

> **说明** 其中${t1}指代的是**SQL脚本-1**节点的输入源的表名称。SQL脚本类的节点支持4个输入源，根据拉线组成上下游关系时，接入SQL脚本节点的接入点位置不同，输入源的表名称依次为t1~t4，您需要根据实际情况调整上述代码中的输入源名称。

v. 单击画布上方**运行**按钮，实验将根据工作流顺序依次运行**读数据表-1**、**SQL脚本-1**节点。

2. 数据预处理：转换所有字段为double。



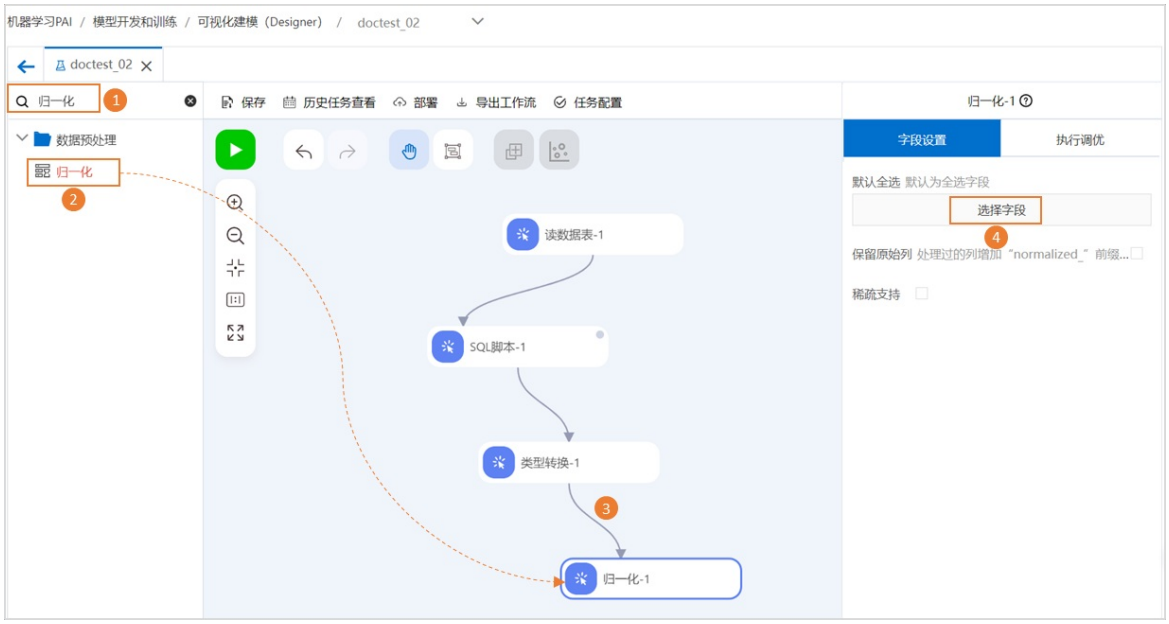
i. 在左侧组件列表的搜索框中，搜索**类型转换**，找到**类型转换**组件。

ii. 将**类型转换**组件拖入右侧画布中，画布中自动生成一个名称为**类型转换-1**的工作流节点。

iii. 通过连线，将**类型转换-1**节点作为**SQL脚本-1**节点的下游节点。

iv. 在画布中选中**类型转换-1**节点，在右侧节点配置页面中**字段设置**页签，单击**转换为double类型的列**下的**选择字段**，将所有字段转换为double类型。

3. 数据预处理：归一化。



i. 在左侧组件列表的搜索框中，搜索**归一化**，找到**归一化**组件。

ii. 将**归一化**组件拖入右侧画布中，画布中自动生成一个名称为**归一化-1**的工作流节点。

iii. 通过连线，将**归一化-1**节点作为**类型转换-1**节点的下游节点。

iv. 在画布中选中**归一化-1**节点，在右侧节点配置页面中的**字段设置**页签，选择所有字段。

4. 在左侧组件列表的搜索框中，搜索**拆分**，找到**拆分**组件后拖入画布，并通过拉线作为**归一化-1**节点的下游节点。

   拆分组件默认将原始数据按4:1拆分为模型训练集和模型预测集。您也可以单击**拆分**组件，在右侧**参数设置**页签，设置**切分比例**。

5. 单击画布上方的**保存**，保存工作流配置。

调试运行工作流

单击画布上方**运行**按钮，调试运行本工作流。

单击运行按钮后，实验将根据工作流顺序依次运行各节点，当节点成功运行完成后，节点的右上角会出现成功提示。运行成功后，您可以右键各节点，选择**查看数据**，进一步查看当前节点的输出数据是否正确。

> **说明** 当工作流比较复杂时，您可以每拖入一个组件配置完成一个节点后，就单击保存和运行，进行调试。如果运行失败可右键节点查看日志，进行失败原因的排查处理。

后续步骤

完成数据预处理后，需要进行数据可视化，详情请参见数据可视化。

2.3. 数据可视化

本文以统计全表信息为例，为您介绍如何进行数据可视化。

前提条件

完成数据准备与预处理，详情请参见数据准备与预处理。

操作步骤

1. 登录PAI控制台，进入工作流页面。

   操作详情请参见step1：进入工作流页面。

2. 构建全表统计节点并运行。

   i. 在组件列表中搜索**全表统计**组件，找到后将其拖入画布，并将生成的节点作为数据准备与预处理中的**类型转换-1**节点的下游节点。

   ii. 右键单击画布中的**全表统计-1**节点，在快捷菜单，单击**执行该节点**。

3. 任务运行结束后，右键单击画布中的**全表统计**组件，在快捷菜单，单击**查看数据**，可以查看数据的全表统计信息。

数据-全表统...

	A	B	C	D	E	F	G	H	I	J
	colname	datatype	totalcount	count	missingcount	nancount	positiveinfinitycour	negativeinfinitycou	min	max
1	age	double	303	303	0	0	0	0	29	77
2	ca	double	303	303	0	0	0	0	0	3
3	chol	double	303	303	0	0	0	0	126	564
4	cp	double	303	303	0	0	0	0	0	2
5	exang	double	303	303	0	0	0	0	0	1
6	fbs	double	303	303	0	0	0	0	0	1
7	oldpeak	double	303	303	0	0	0	0	0	6.2
8	restecg	double	303	303	0	0	0	0	0	2
9	sex	double	303	303	0	0	0	0	0	1
10	slop	double	303	303	0	0	0	0	0	2
11	status	double	303	303	0	0	0	0	0	1
12	style	double	303	303	0	0	0	0	0	3
13	thal	double	303	303	0	0	0	0	0	2
14	thalach	double	303	303	0	0	0	0	71	202
15	trestbps	double	303	303	0	0	0	0	94	200

后续步骤

完成数据可视化后，可以进行算法建模，详情请参见算法建模。

2.4. 算法建模

本文以逻辑回归二分类算法为例，为您介绍如何使用PAI Designer训练模型。

前提条件

完成数据可视化，详情请参见数据可视化。

操作步骤

1. 登录PAI控制台，进入工作流页面。

   操作详情请参见step1：进入工作流页面。

2. 构建逻辑回归二分类节点并运行。

   i. 在组件列表中搜索**逻辑回归二分类**组件，找到后将其拖入画布，并将生成的节点作为数据准备与预处理中的**拆分-1**节点的下游节点。

   ii. 单击画布中的**逻辑回归二分类-1**节点，在右侧**字段设置**页签，将**目标列**设置为status，将**训练特征列**设置为除**目标列**以外的所有列。

   iii. 右键单击画布中的**逻辑回归二分类-1**节点，在快捷菜单，单击**执行该节点**。

后续步骤

完成算法建模后，可以对训练模型进行评估，详情请参见模型评估。

2.5. 评估模型

本文以评估二分类训练模型为例，为您介绍如何进行模型评估。

前提条件

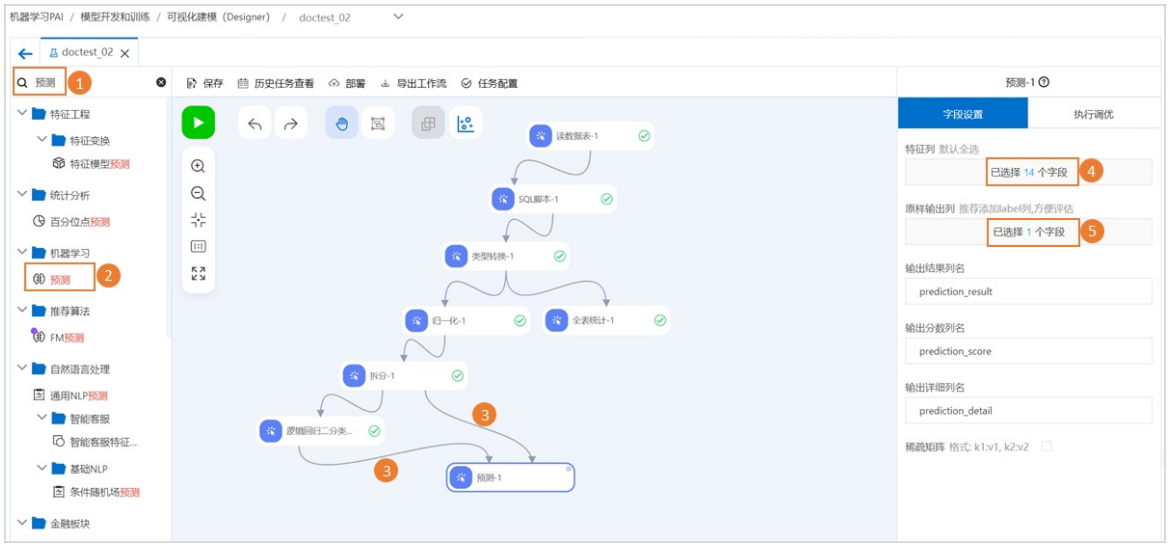
完成建模，详情请参见算法建模。

操作步骤

1. 登录PAI控制台，进入工作流页面。

   操作详情请参见step1：进入工作流页面。

2. 构建预测节点。

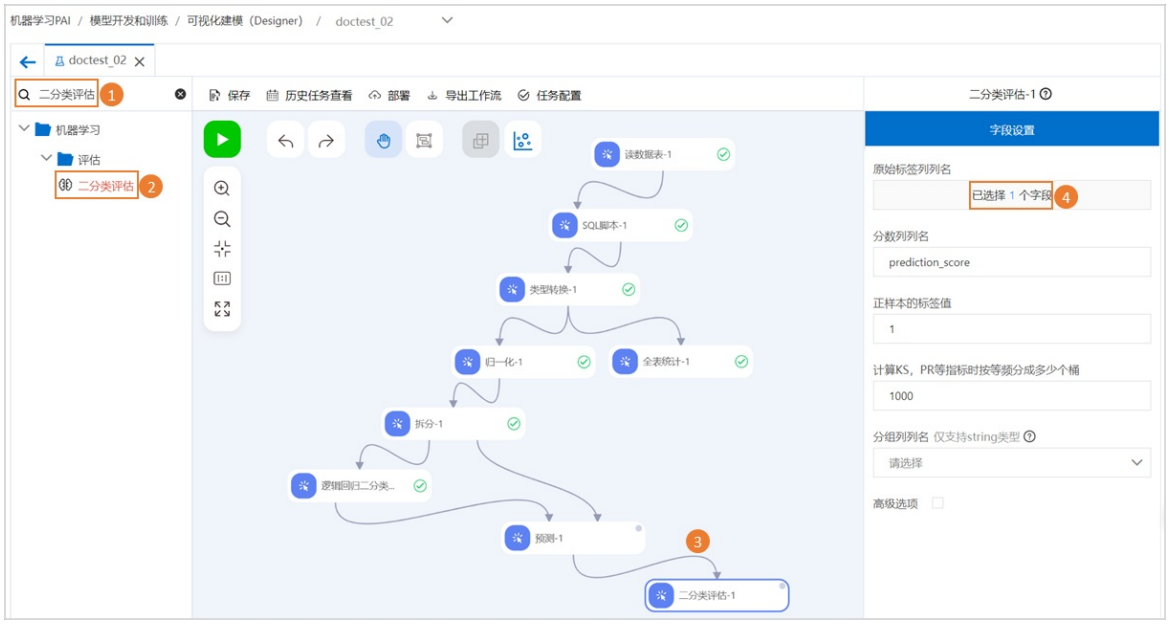


i. 在组件列表中分别搜索**预测**组件，找到后将其拖入画布，并将生成的节点作为**拆分-1**、**逻辑回归二分类-1**节点的下游节点，拼接为实验。

ii. 单击画布中的**预测-1**节点，在右侧节点配置中，分别单击**特征列字段**、**原样输出列字段**下的**选择字段**，在**已选**模块切换至**编辑**状态进行如下配置。

- 特征列字段：在编辑框中输入除 `status` 外的其他所有字段：`age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slop,ca,thal,style`。
- 原样输出列字段：在编辑框中输入 `status` 字段。

3. 构建二分类评估节点。



i. 在组件列表中分别搜索**二分类评估**组件，找到后将其拖入画布，并将生成的节点作为**预测-1**节点的下游节点，拼接为实验。

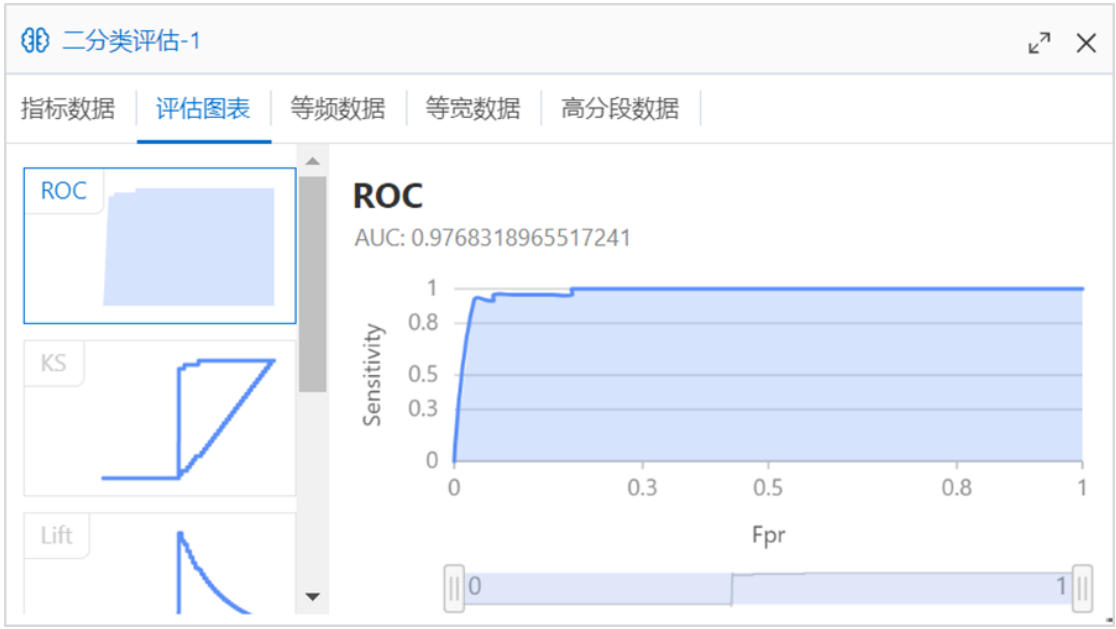
ii. 单击画布中的**二分类评估-1**节点，在右侧**字段设置**页签，将**原始标签列列名**设置为status。

4. 单击画布上方的**运行**。

5. 查看模型评估结果。

i. 实验运行结束后，右键单击画布中的**二分类评估**组件，在快捷菜单，单击**可视化分析**。

ii. 单击**评估图表**页签，查看不同参数情况下，二分类训练模型的ROC（接受者操作特性）曲线。



3.模板工作流demo

PAI-Designer为您提供了丰富的工作流模板，您可以直接套用模板快速构建模型。本文使用**心脏病预测案例**模板为例，为您介绍如何快速地构建工作流。

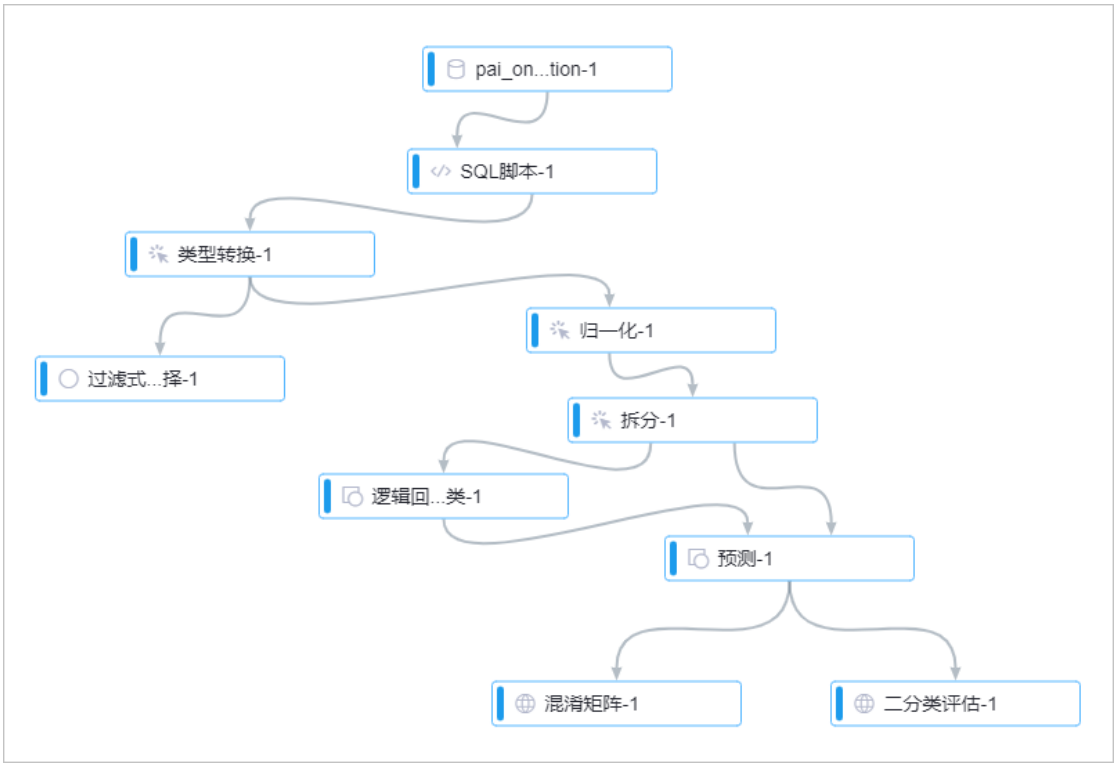
前提条件

已开通PAI并创建了工作空间。

操作步骤

1. 进入PAI-Designer页面。
   i. 登录PAI控制台。
   ii. 在左侧导航栏单击**工作空间列表**，在工作空间列表页面中单击待操作的工作空间名称，进入对应工作空间内。
   iii. 在工作空间页面的左侧导航栏选择**模型开发和训练 > 可视化建模（PAI-Designer）**，进入Designer页面。
2. 构建并运行实验。
   i. 在Designer页面单击**工作流模板**，在全部模板页签内找到**心脏病预测案例**。
   ii. 单击**心脏病预测案例**下的**创建**，在弹出的页面中单击**确定**。

   等待大约十秒钟，实验构建成功后如下图所示。



   iii. 单击画布上方的**运行**，工作流中的节点运行完成后，可右键单击对应节点，查看节点的输出。