用户指南

# 通用约定

| 格式 | 说明 | 样例 |
| --- | --- | --- |
|  | 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。 |
|  | 该类警示信息可能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，恢复业务所需时间约10分钟。 |
|  | 用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。 | **注:**您也可以通过按**Ctrl** + **A**选中全部文件。 |
| > | 多级菜单递进。 | **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。 | 单击 **确定**。 |
| courier字体 | 命令。 | 执行 cd /d C:/windows 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。 | swich {stand | slave} |

内容

通用约定 1

1机器学习平台 3

1.1机器学习的一般过程 3

1.1.1 新建实验 3

1.1.2读取数据表 3

1.1.2.1 查看字段信息 4

1.1.3数据预处理 4

1.1.3.1缺失值填充 4

1.1.3.2缺失值填充－参数设置 5

1.1.3.3数据拆分 5

1.1.4数据探索 6

1.1.4.1查看全表统计信息 6

1.1.5算法建模 7

1.1.5.1 选择算法组件 7

1.1.5.2 算法参数配置 8

1.1.5.3 预测和评估的参数配置 8

1.1.6 模型评估 9

1.1.6.1 查看模型 9

1.1.6.2 二分类评估 10

1.1.7 数据存储 11

1.1.7.1 写数据表 11

1.1.7.2 保存模型 12

# 1机器学习平台

## 1.1机器学习的一般过程

一个完整的建模步骤包括以下6个步骤：

step1：数据准备

step2：数据预处理

step3：数据可视化

step4：算法建模

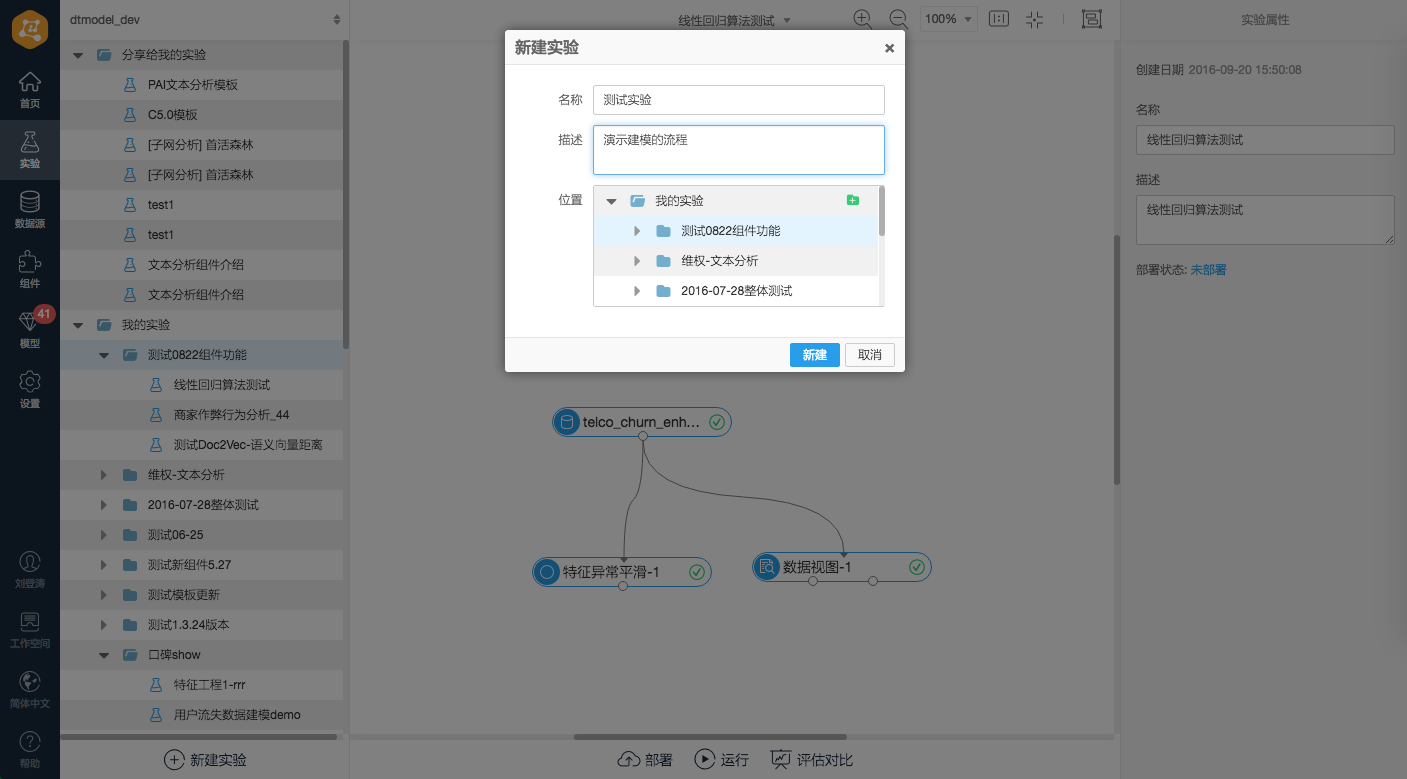
step5：模型评估

step6：数据存储

注意：以下步骤默认用户已拥有自己的projet并将表数据传入对应的project中

### 1.1.1 新建实验

新建空白实验，输入实验名和实验描述:

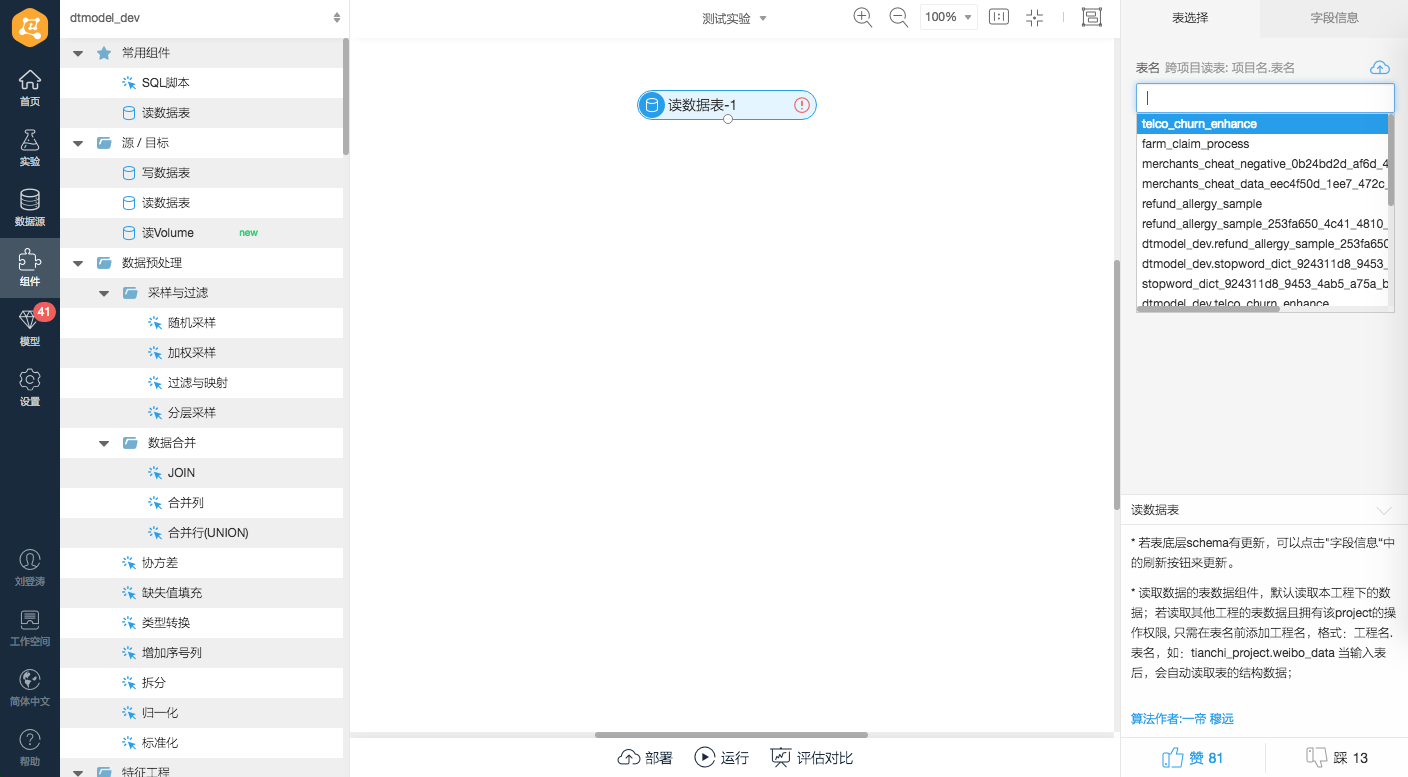


其中：

* 名称：对创建的实验命名。
* 描述：对实验内容的描述。
* 位置：实验创建后放置的位置。

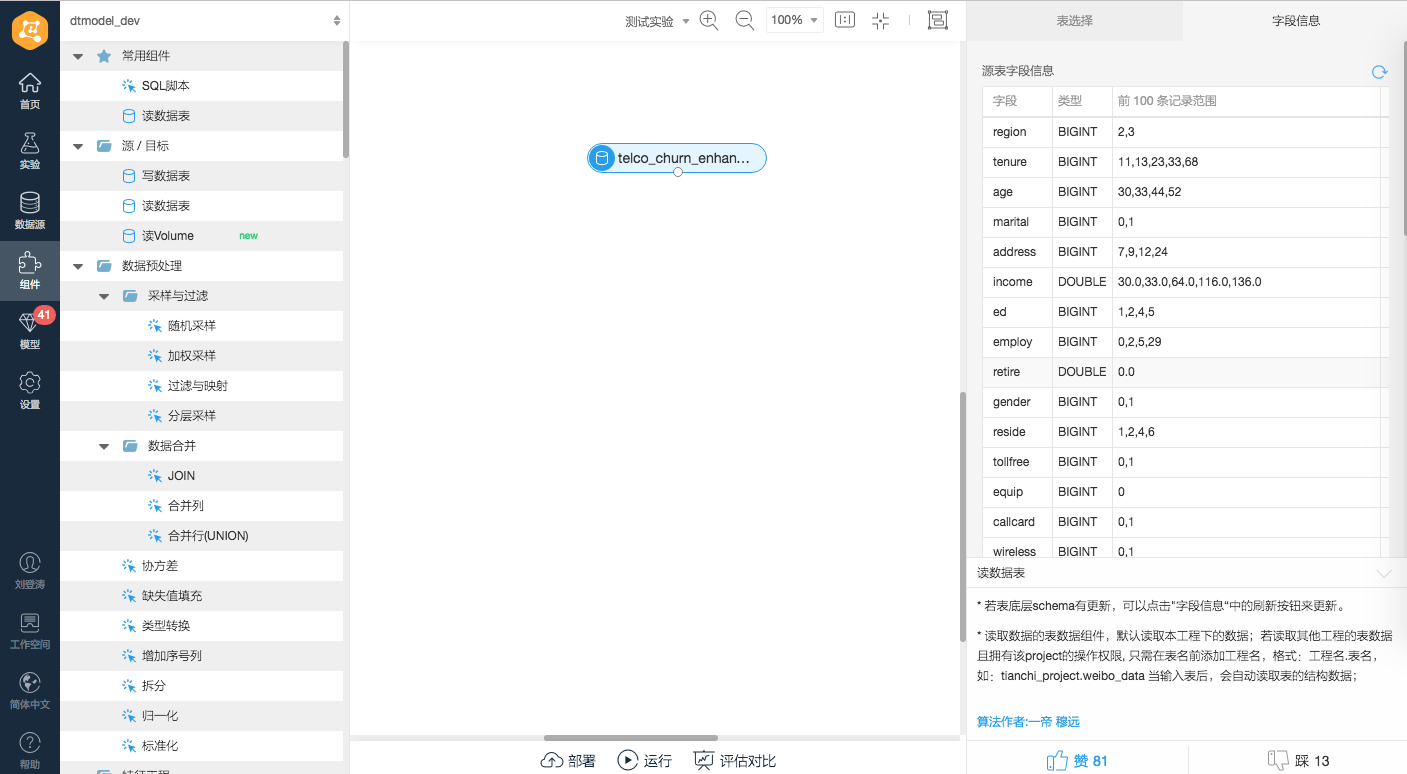
### 1.1.2读取数据表

从组件栏中拖入读数据表，配置你的ODPS表：



### 1.1.2.1 查看字段信息

* 切换到字段信息栏，可以查看输入表的字段名、数据类型和前100行数据的数值分布。
* 在画布中右键点击组件可查看前100条详细数据。

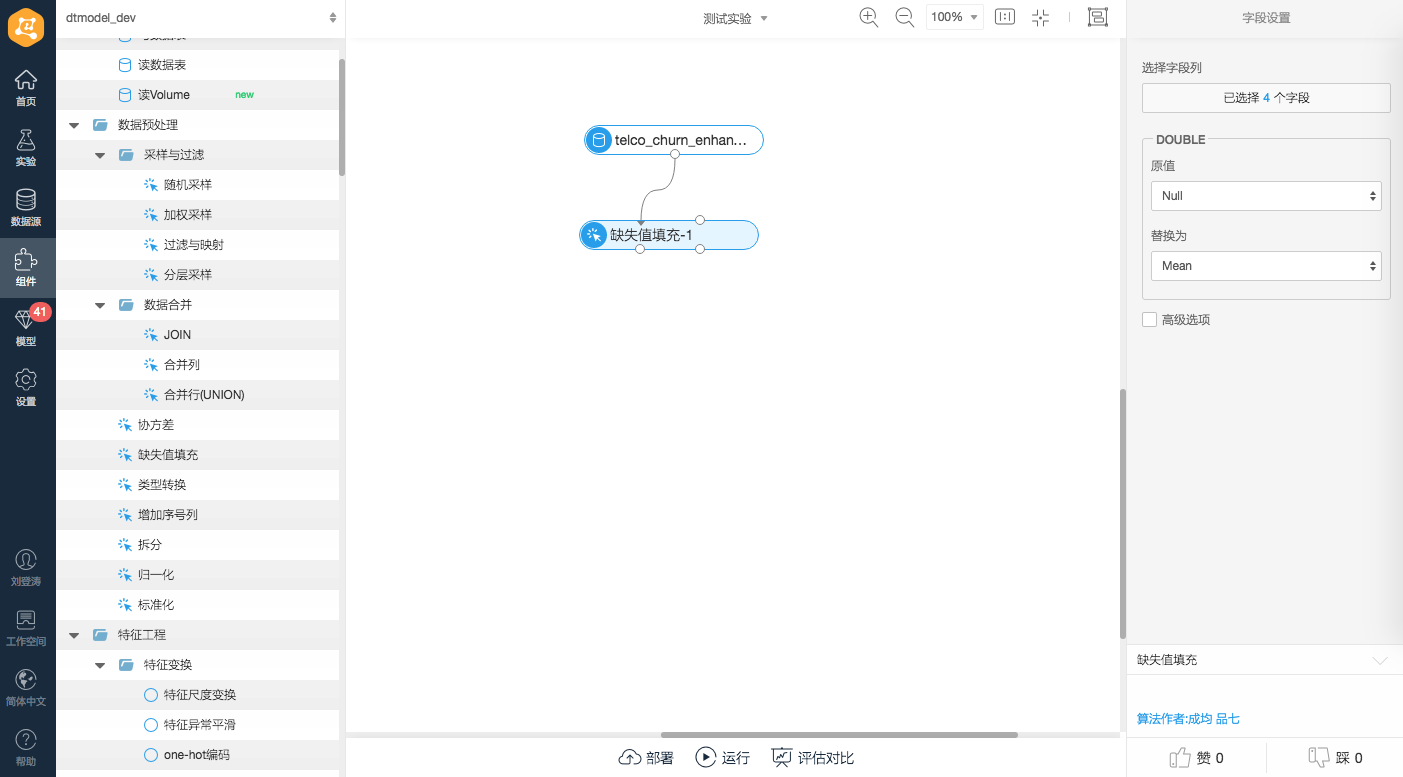


### 1.1.3数据预处理

对前一步导入的数据进行预处理。

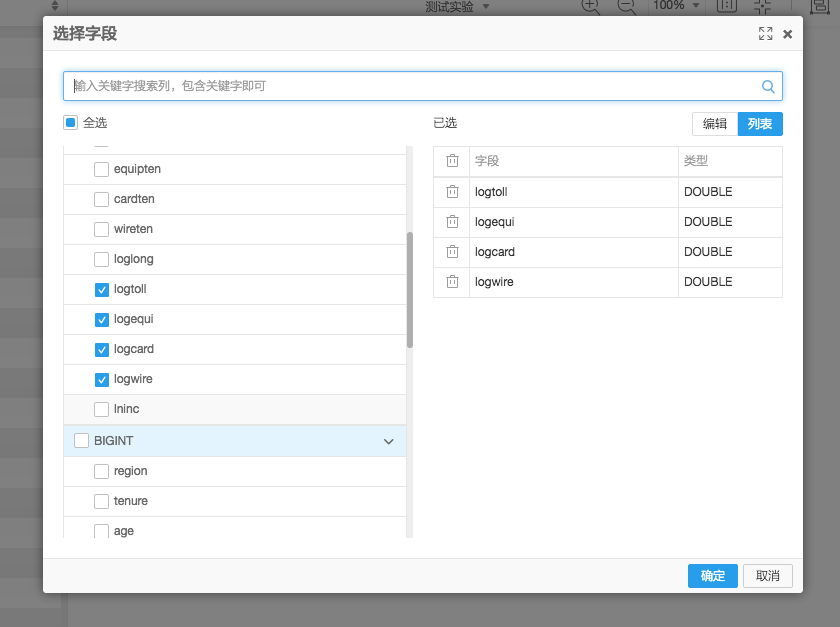
### 1.1.3.1缺失值填充

拖入“数据预处理–缺失值填充”组件，并将两个组件连线



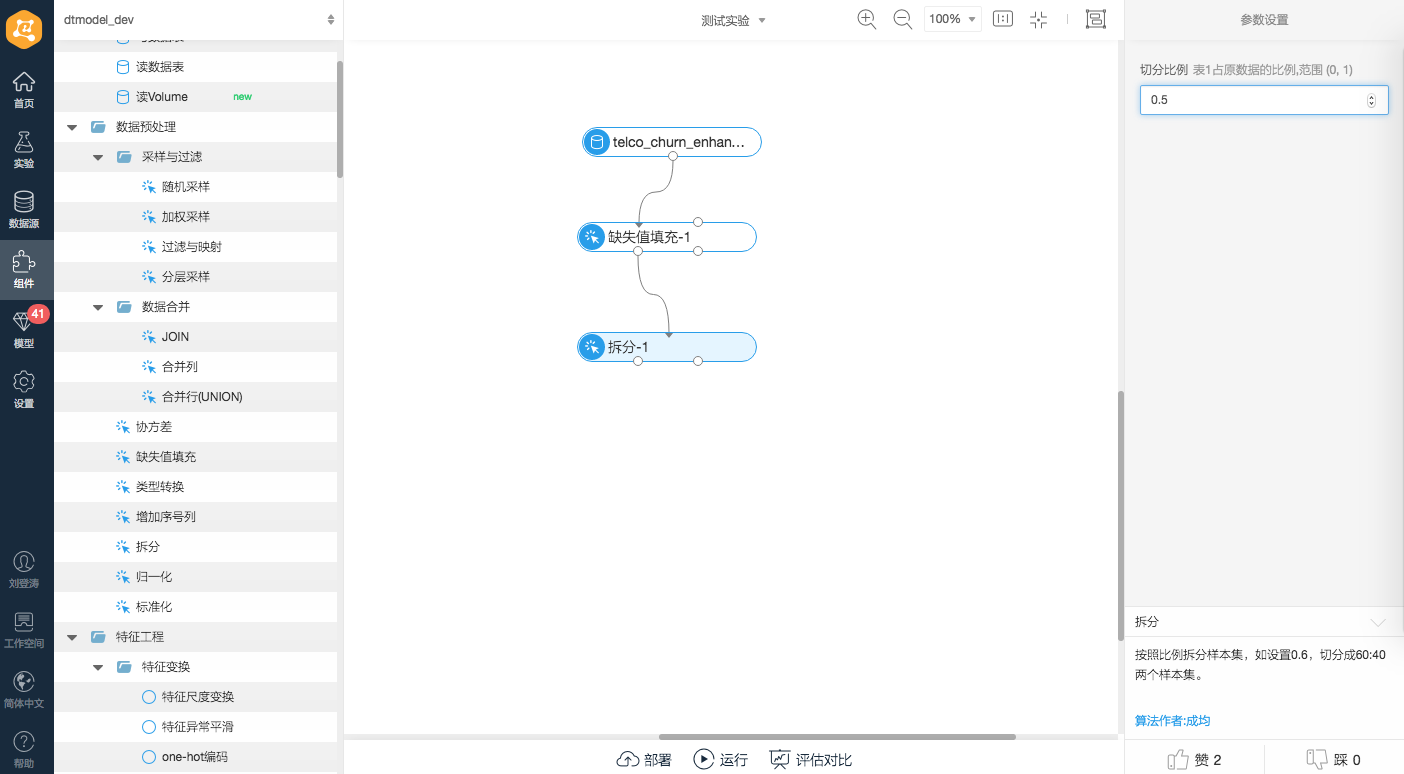
### 1.1.3.2缺失值填充－参数设置

设置组件中需要处理的字段和填充方法



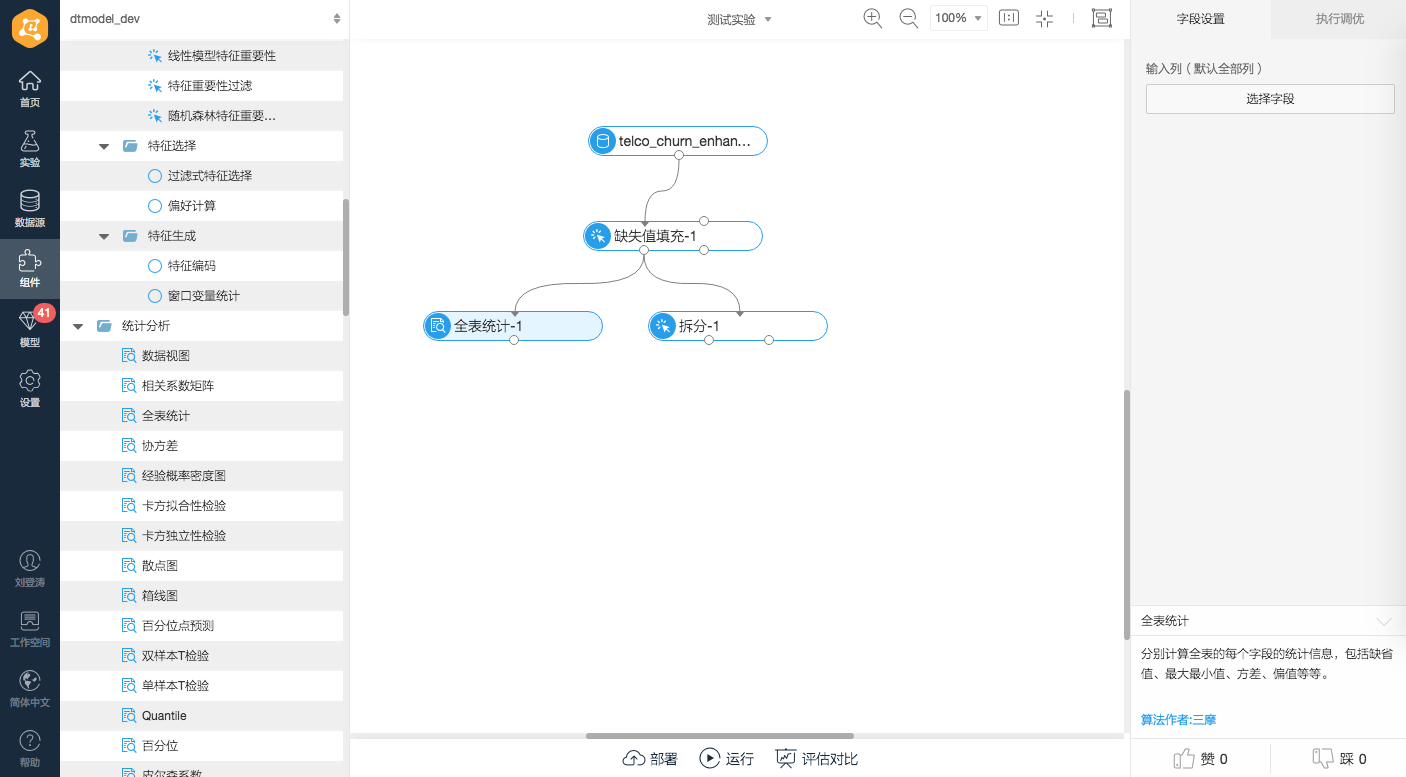
### 1.1.3.3数据拆分

* 对处理好的数据进行拆分。
* 拖入拆分组件，用线条连接缺失值填充和拆分组件。
* 此步骤的目的是将数据拆分成两份，例如：50%作为模型训练集，50%作为模型预测集。



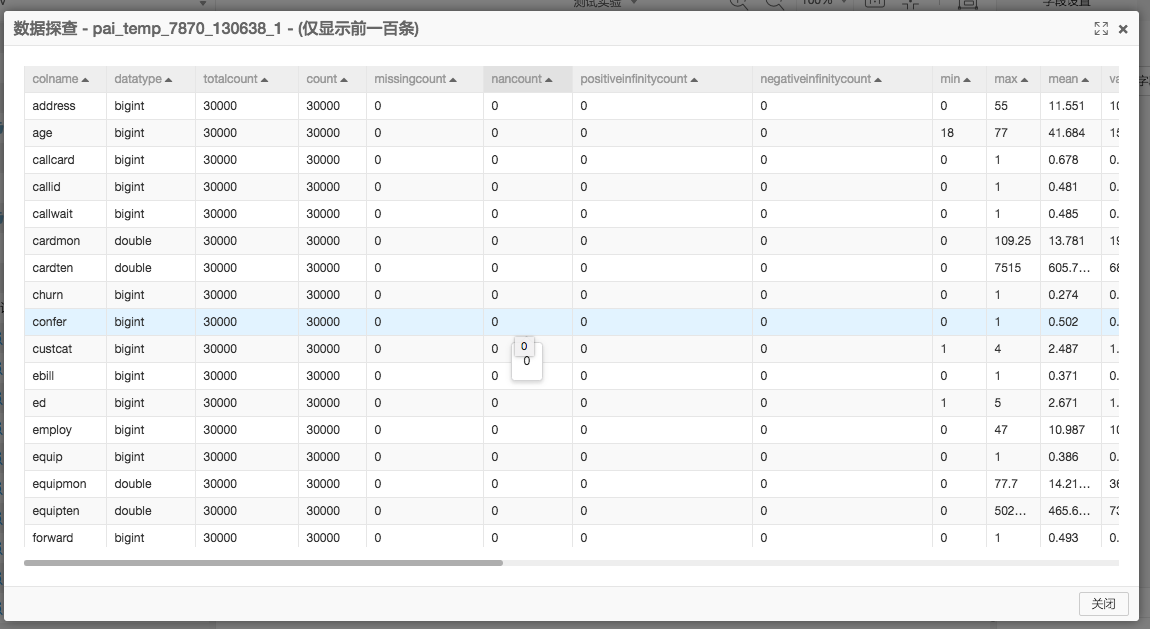
### 1.1.4数据探索

* 用“全表统计”对数据做一个简单的分析，使用搜索或从“组件－数据探索－统计分析－基本分析－全表统计”拖入画布中。
* 将“缺失值填充”和“全表统计”两个组件用线条连接。



### 1.1.4.1查看全表统计信息

点击"运行"后，右键点击“全表统计”选择查看分析报告，可看到数据的全表统计信息，如下：

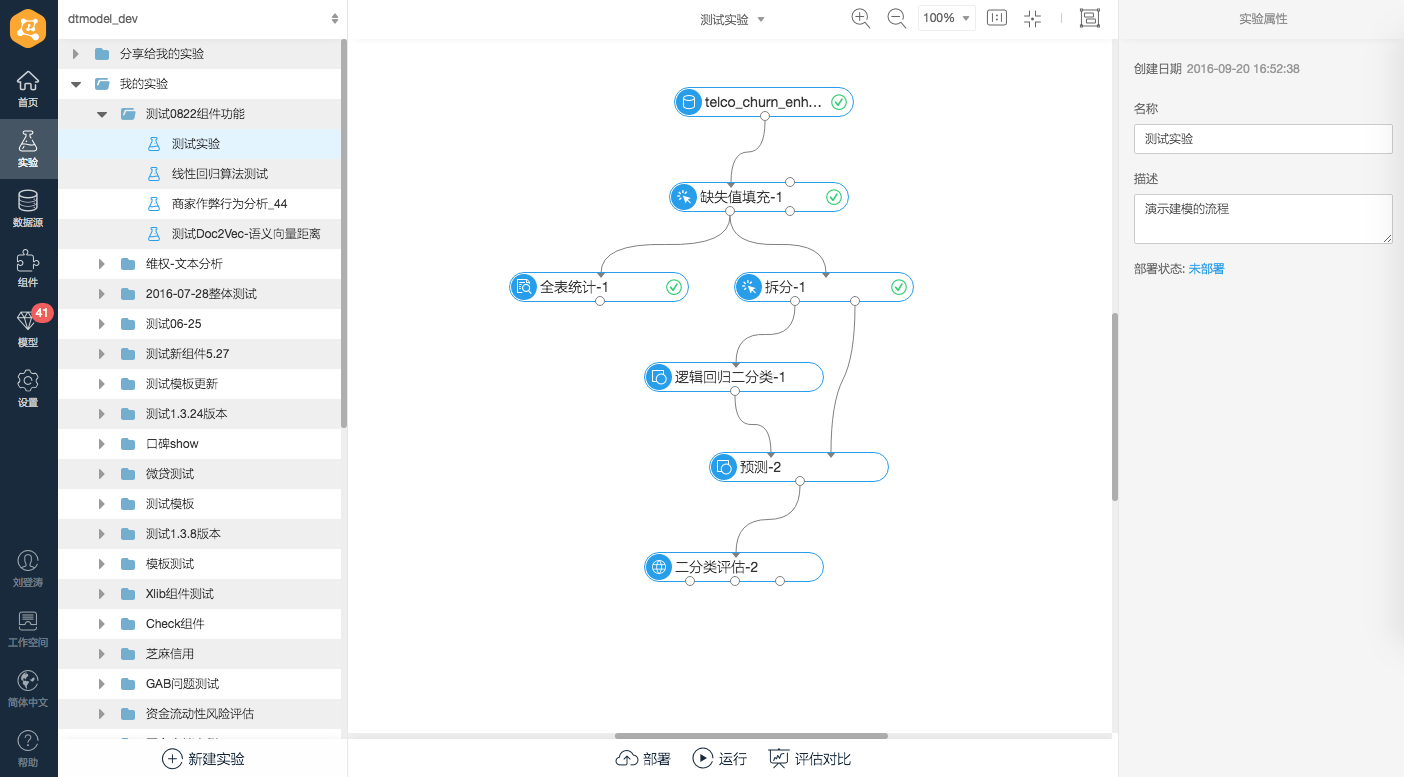


### 1.1.5算法建模

### 1.1.5.1 选择算法组件

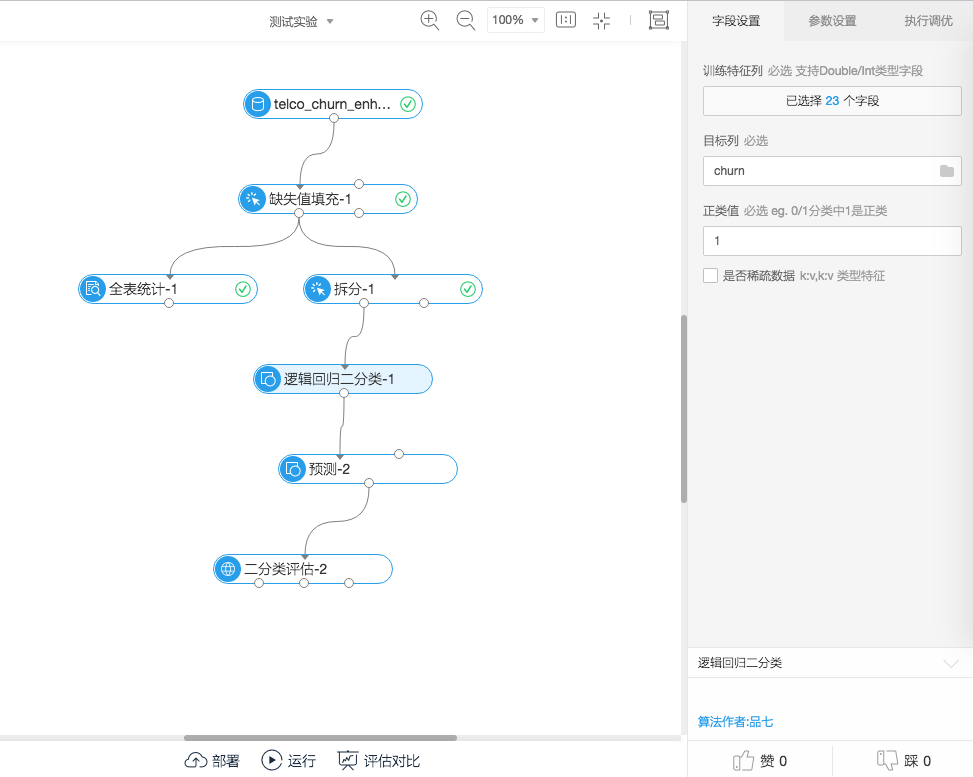
本实验使用二分类模型，分别从“机器学习”栏目，拖入–逻辑回归二分类、预测、二分类评估组件，

将组件用线条连接，如下图：



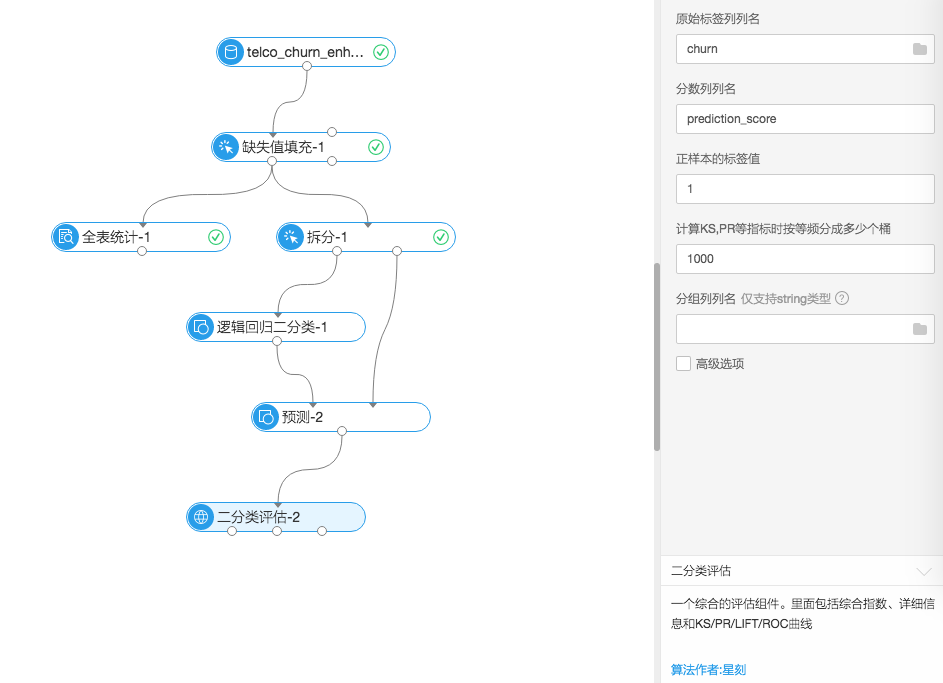
### 1.1.5.2 算法参数配置

选中“逻辑回归二分类”组件，画布右侧出现逻辑回归参数配置字段–训练特征列和目标列，如下图：



### 1.1.5.3 预测和评估的参数配置

预测组件第一个输入为算法模板，第二个输入为测试集，字段和参数无需配置；二分类评估组件参数“原始标签列列名”与逻辑回归组件“目标列”参数配置成相同的字段，如下：



预测组件输出字段解释

predict\_result:预测结果列；

predict\_score：预测结果概率得分； 仅模型为二分类时有效

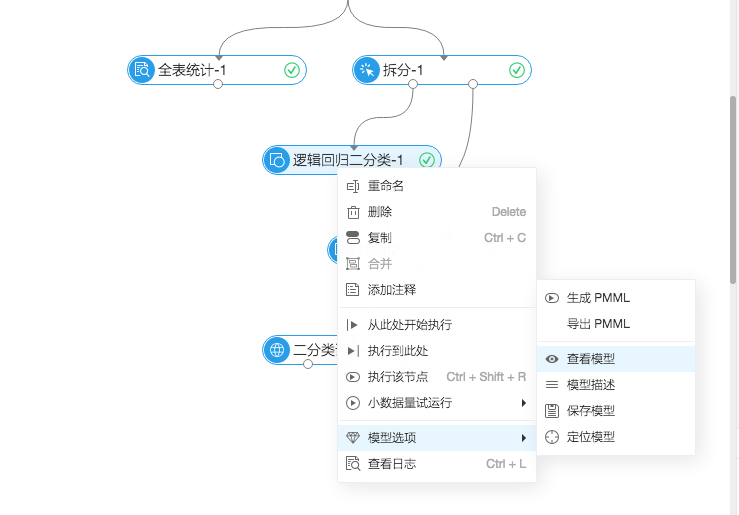
predict\_detail：每个类别的预测概率得分；仅模型为二分类时有效

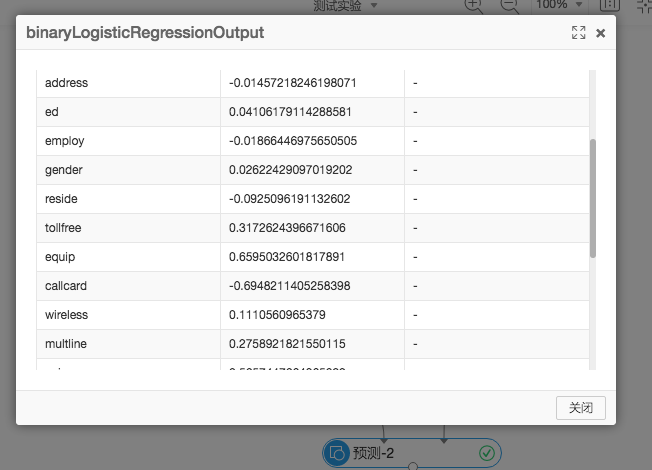
### 1.1.6 模型评估

运行以上流程，可查看模型的信息。

### 1.1.6.1 查看模型

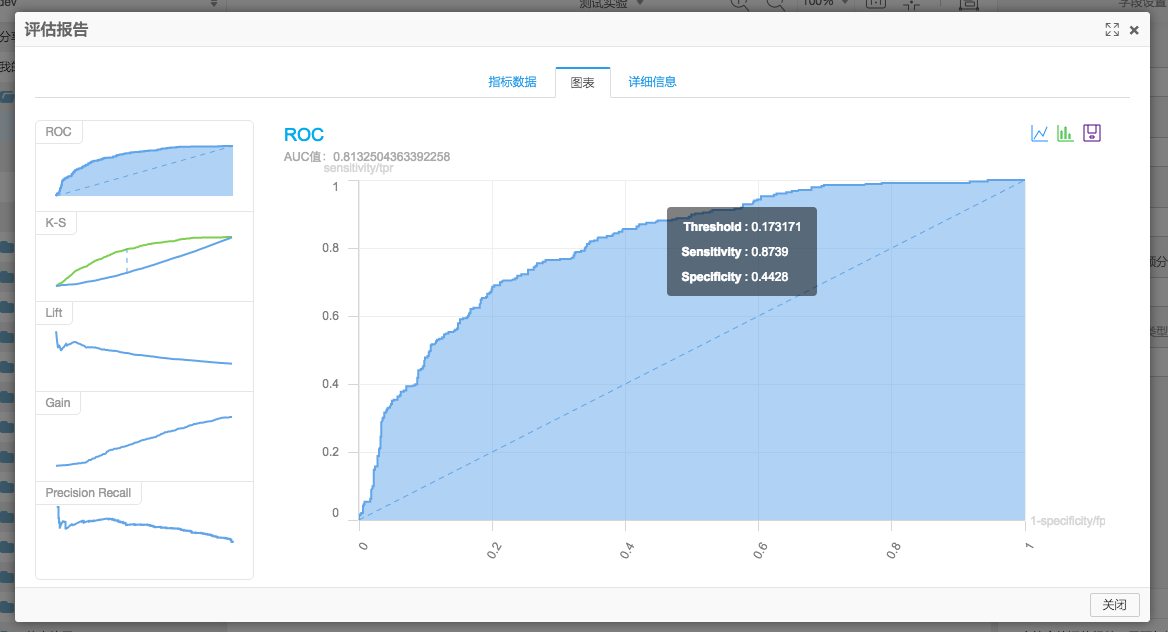
右键点击“逻辑回归二分类”，点击“模型选项－查看模型”，如下图：





### 1.1.6.2 二分类评估

右键点击“二分类评估”，点击“查看评估报告”，如下图：

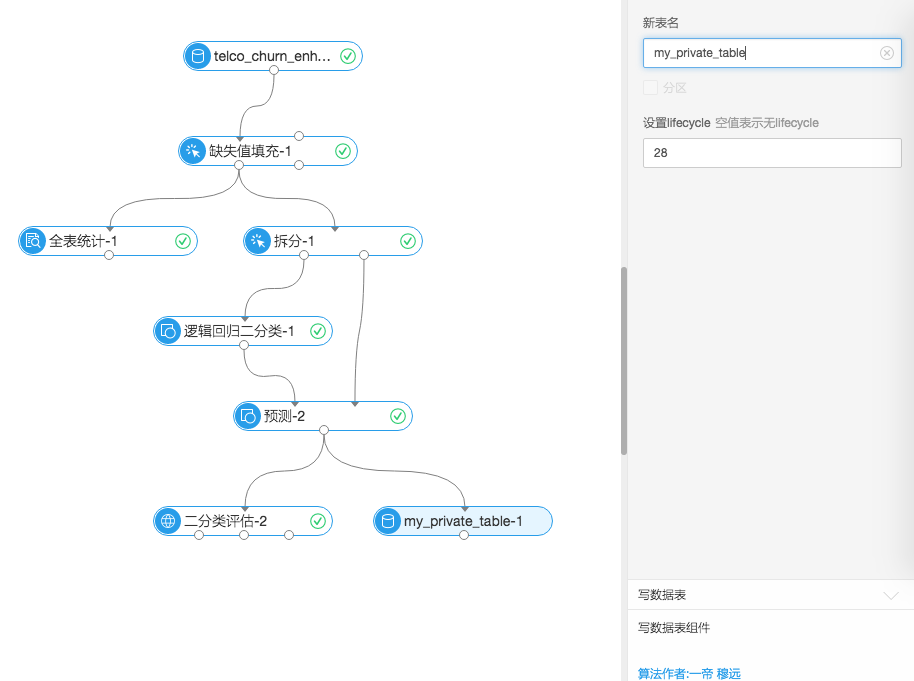


### 1.1.7 数据存储

### 1.1.7.1 写数据表

搜索“写数据表”组件，或者从“组件－数据读写－写数据表”中选中组件拖入画布中。

流程中任意节点的中间数据存储指定的ODPS表，如下图：



### 1.1.7.2 保存模型

选中训练好的模型，右键点击，选择“模型选项－保存模型”，保存后的模型可以在左侧的“模型”一级菜单中找到。

选择训练好的模型可以拖入到画布中，可实现模型的预测功能，如下图：

