




# 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。	<b>危险:</b> 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。	<b>警告:</b> 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。	<b>注:</b> 您也可以通过按 <b>Ctrl + A</b> 选中全部文件。
>	多级菜单递进。	<b>设置 &gt; 网络 &gt; 设置网络类型</b>
<b>粗体</b>	表示按键、菜单、页面名称等 UI 元素。	单击 <b>确定</b> 。
courier 字体	命令。	执行 <code>cd /d C:/windows</code> 命令，进入 Windows 系统文件夹。
<i>斜体</i>	表示参数、变量。	<code>bae log list --instanceid <i>Instance_ID</i></code>
[ ] 或者 [a b]	表示可选项，至多选择一个。	<code>ipconfig [-all/-t]</code>
{ } 或者 {a b}	表示必选项，至多选择一个。	<code>swich {stand / slave}</code>

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

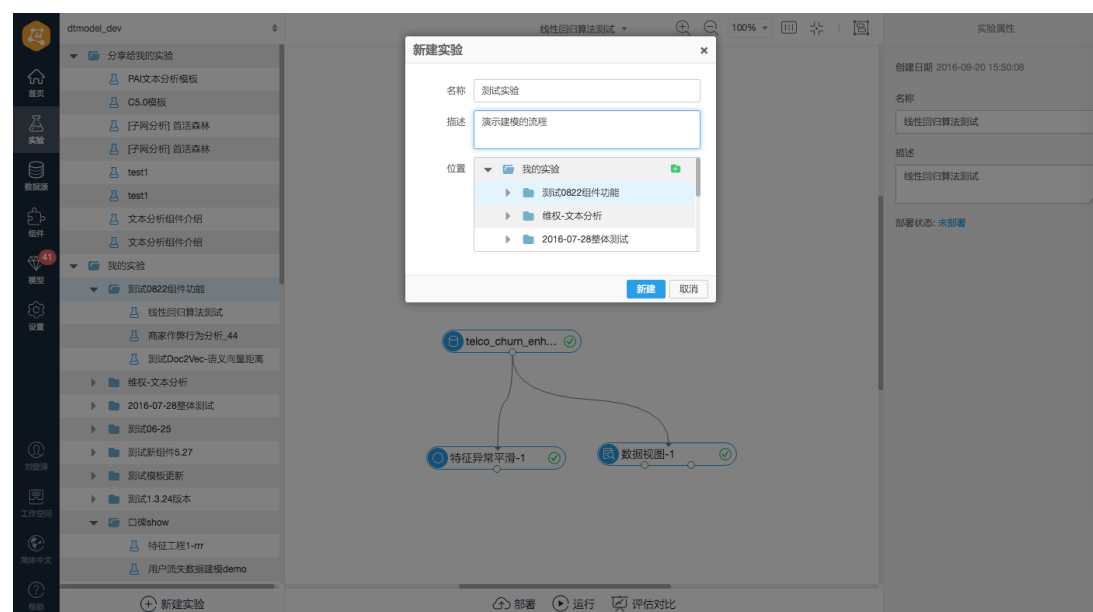
step5: 模型评估

step6: 数据存储

注意：以下步骤默认用户已拥有自己的 project 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：

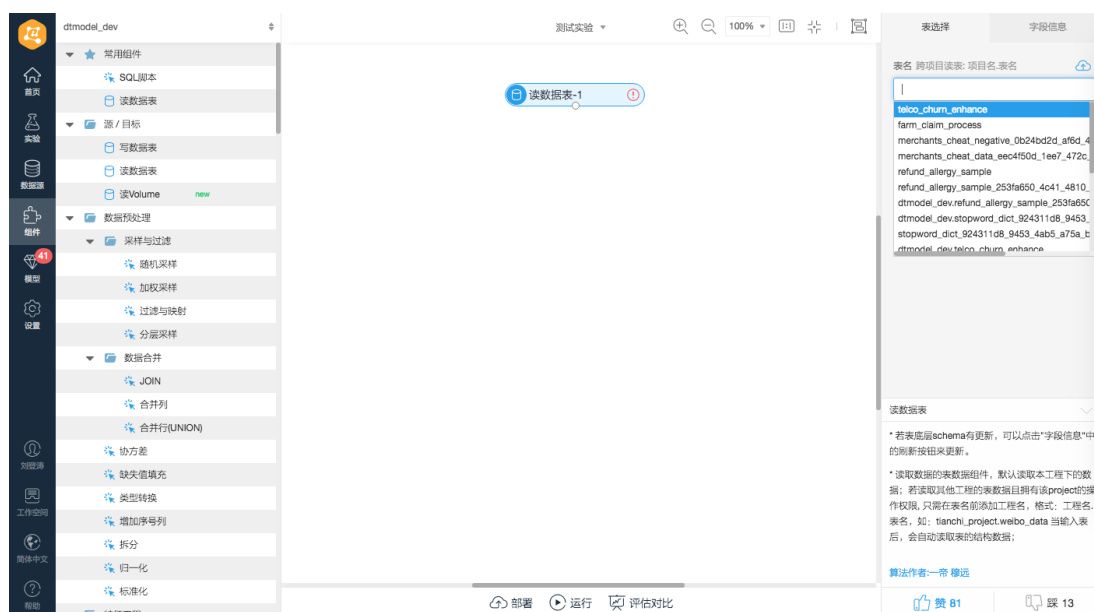


其中：

- 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。

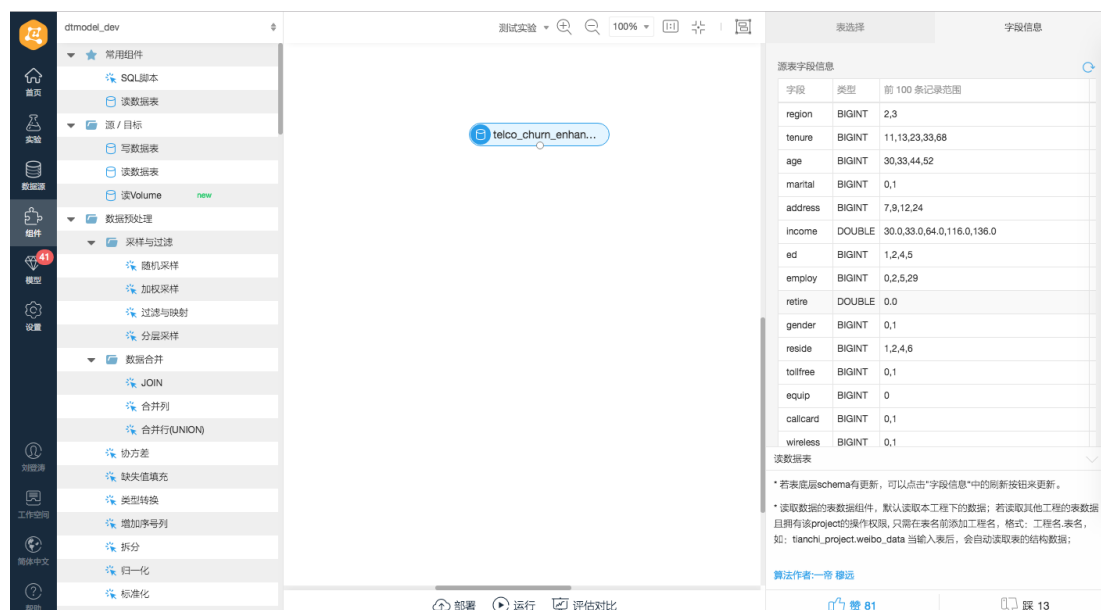
### 1.1.2 读取数据表

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。

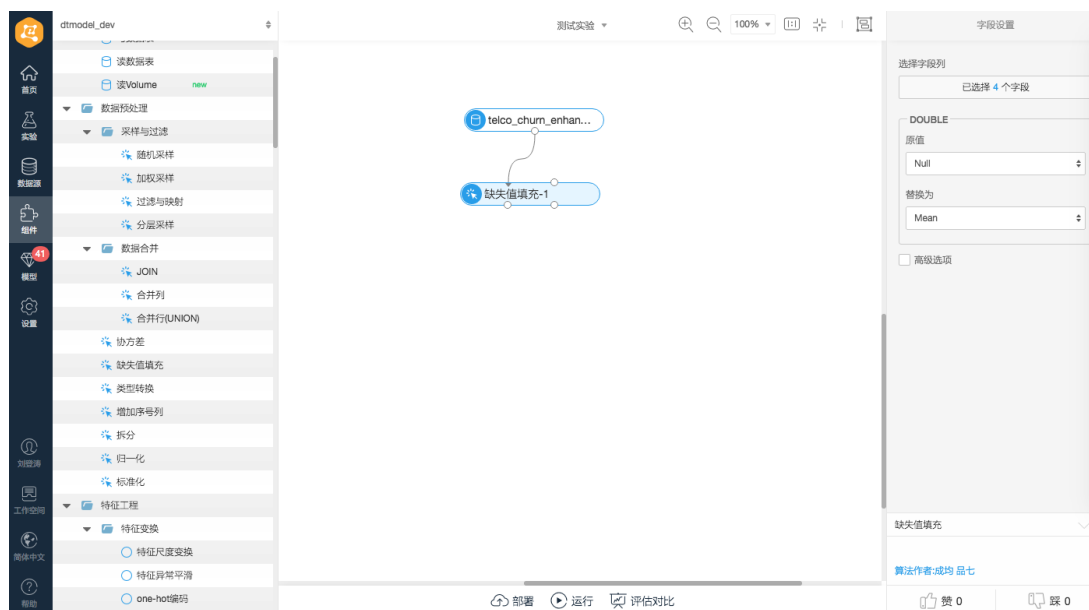


## 1.1.3 数据预处理

对前一步导入的数据进行预处理。

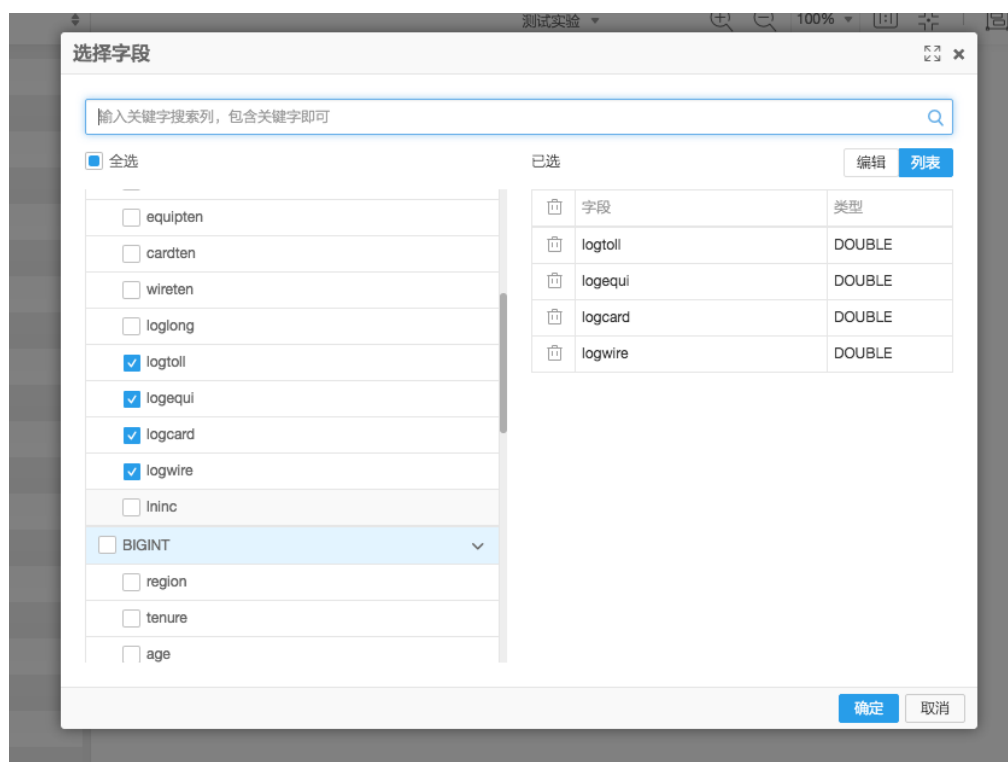
### 1.1.3.1 缺失值填充

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

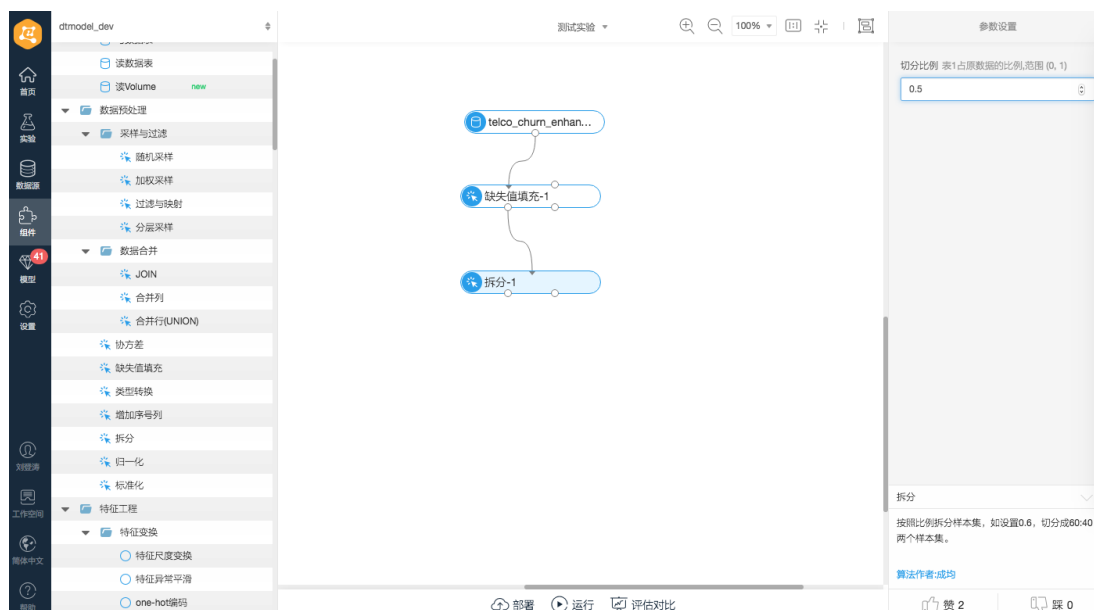
设置组件中需要处理的字段和填充方法



## 1.1.3.3 数据拆分

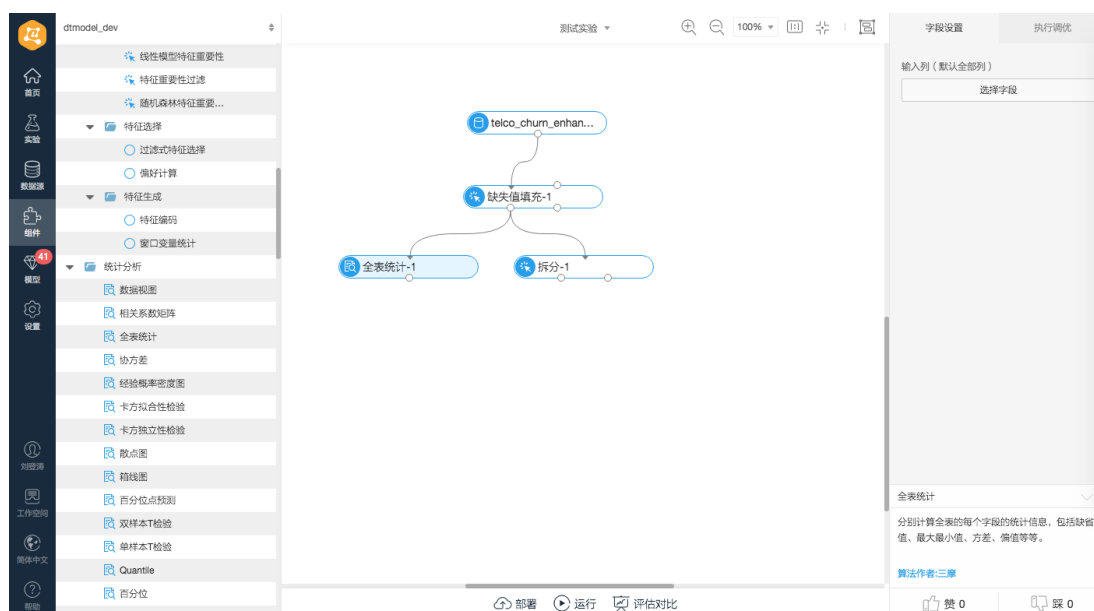
- 对处理好的数据进行拆分。
- 拖入拆分组件，用线条连接缺失值填充和拆分组件。

- 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：

数据探查 - pai\_temp\_7870\_130638\_1 - (仅显示前一百条)

colname	datatype	totalcount	count	missingcount	nancount	positiveinfinitycount	negativeinfinitycount	min	max	mean	var
address	bigint	30000	30000	0	0	0	0	0	55	11.551	11.551
age	bigint	30000	30000	0	0	0	0	18	77	41.684	11.551
callcard	bigint	30000	30000	0	0	0	0	0	1	0.678	0.000
callid	bigint	30000	30000	0	0	0	0	0	1	0.481	0.000
callwait	bigint	30000	30000	0	0	0	0	0	1	0.485	0.000
cardmon	double	30000	30000	0	0	0	0	0	109.25	13.781	11.551
cardten	double	30000	30000	0	0	0	0	0	7515	605.7...	605.7...
churn	bigint	30000	30000	0	0	0	0	0	1	0.274	0.000
confer	bigint	30000	30000	0	0	0	0	0	1	0.502	0.000
custcat	bigint	30000	30000	0	0	0	0	1	4	2.487	1.000
ebill	bigint	30000	30000	0	0	0	0	0	1	0.371	0.000
ed	bigint	30000	30000	0	0	0	0	1	5	2.671	1.000
employ	bigint	30000	30000	0	0	0	0	0	47	10.987	11.551
equip	bigint	30000	30000	0	0	0	0	0	1	0.386	0.000
equipmon	double	30000	30000	0	0	0	0	0	77.7	14.21...	14.21...
equipten	double	30000	30000	0	0	0	0	0	502...	465.6...	465.6...
forward	bigint	30000	30000	0	0	0	0	0	1	0.493	0.000

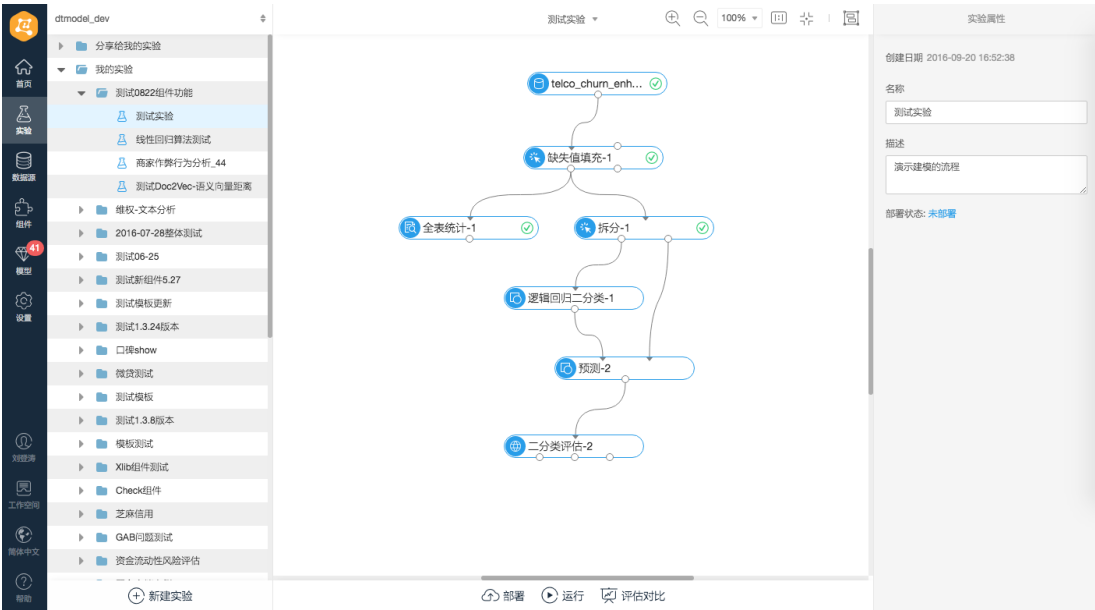
关闭

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

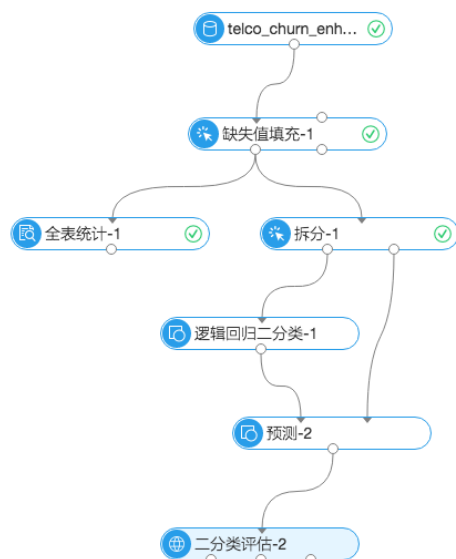
选中“逻辑回归二分类”组件，画布右侧出现逻辑回归参数配置字段 - 训练特征列和目标列，如下图：

The screenshot displays a machine learning workflow in a software interface. The workflow consists of several components connected in a sequence: 'telco\_churn\_enh...' (with a green checkmark), '缺失值填充-1' (Missing Value Imputation-1, with a green checkmark), '全表统计-1' (Full Table Statistics-1, with a green checkmark), '拆分-1' (Split-1, with a green checkmark), '逻辑回归二分类-1' (Logic Regression Binary Classification-1), '预测-2' (Prediction-2), and '二分类评估-2' (Binary Classification Evaluation-2). The '逻辑回归二分类-1' component is currently selected, and its configuration panel is visible on the right. This panel includes tabs for '字段设置' (Field Settings), '参数设置' (Parameter Settings), and '执行调优' (Execution Tuning). Under '参数设置', the '训练特征列' (Training Features) section shows '已选择 23 个字段' (23 fields selected). The '目标列' (Target Column) is set to 'churn'. The '正类值' (Positive Class Value) is set to '1'. There is also a checkbox for '是否稀疏数据' (Whether sparse data) which is currently unchecked. At the bottom of the interface, there are buttons for '部署' (Deploy), '运行' (Run), and '评估对比' (Evaluate and Compare).

### 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：





原始标签列名  
churn

分数列名  
prediction\_score

正样本的标签值  
1

计算KS,PR等指标时按等频分成多少个桶  
1000

分组列名 仅支持string类型

☐ 高级选项

二分类评估  
一个综合的评估组件。里面包括综合指数、详细信息和KS/PR/LIFT/ROC曲线

算法作者:星刻

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

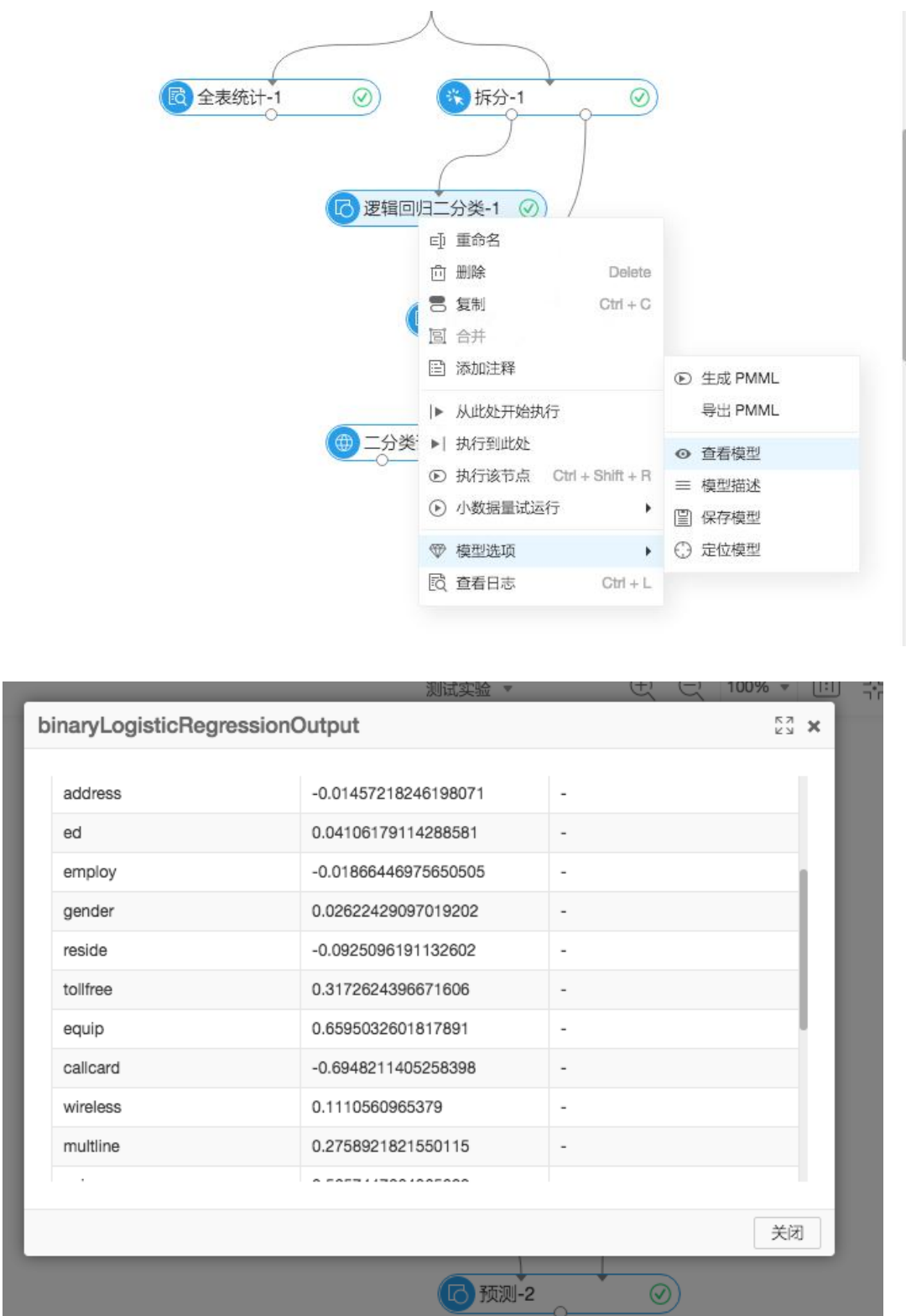
predict\_detail: 每个类别的预测概率得分; 仅模型为二分类时有效

### 1.1.6 模型评估

运行以上流程, 可查看模型的信息。

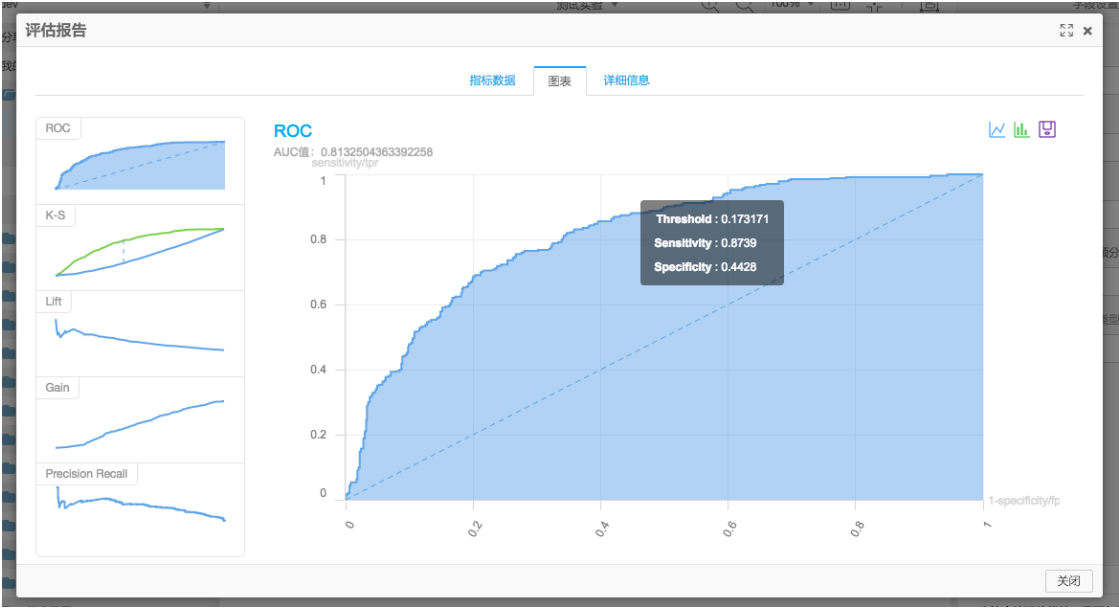
#### 1.1.6.1 查看模型

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：

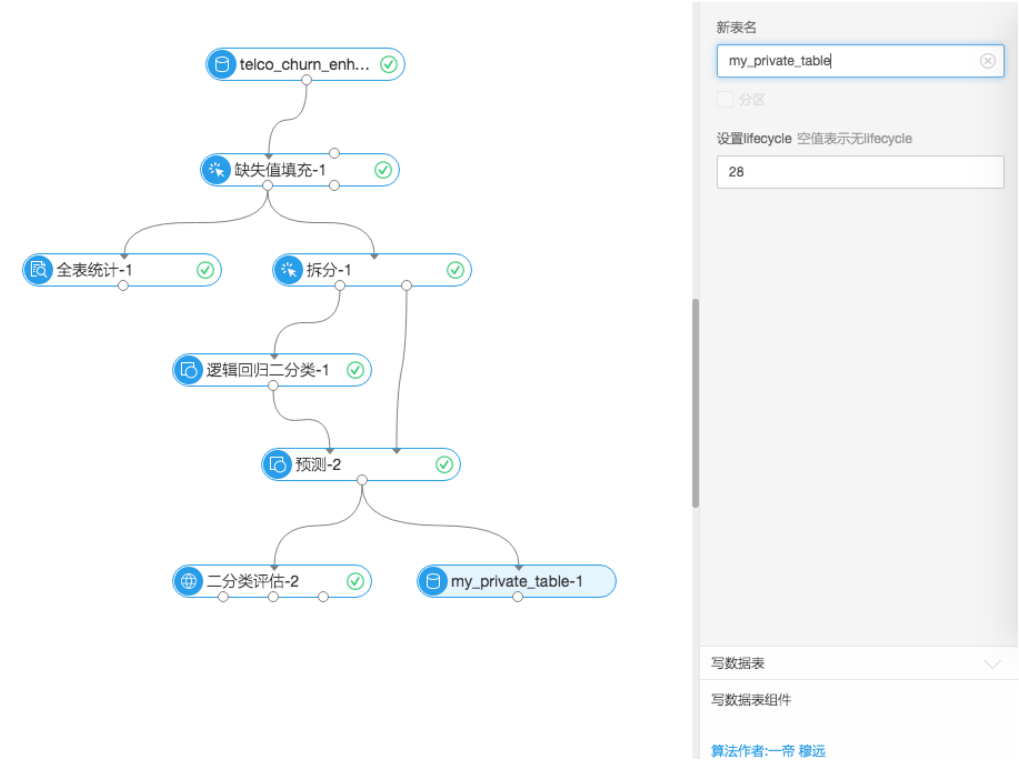


### 1.1.7 数据存储

#### 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：

