1.背景介绍

以客户为中心目标是了解不同客户群体的需求，并向他们提供满足其个性化需求的服务——提供了更多附加值，与此同时也获得高于平均的客户价值。 随着电信业产品同质化的趋势越来越明显，企业成功不能再单纯依靠创新产品和产品价格取胜。倾听客户呼声和需求、对不断变化的客户期望迅速作出反应的以客户为中心运营能力——已经成为当今企业能否成功的关键。

某电信运营商上个月发展了19999个新用户，在新用户入网一个月后，打算做两件事情。

1、希望通过提供一些优惠提高用户的忠诚度

2、希望通过推荐一些产品提升客单价

         为达到这一目的，运营商请数据分析师对新用户进行分析，希望通过对新用户的洞察，搞清楚以下几个问题：

         1）、该给客户提供什么优惠？

              我们的优惠能否给客户带来惊喜？

             不同的客户是否该根据他们的喜好提供不同的优惠？

         2）、客户对我们的什么产品感兴趣？

             不同的客户是否应该推荐不同的产品？

1. 实验思路

数据分析师接到任务后，开始对任务进行梳理，经过分解和排序，分析师给出的解决该问题的思路如下：

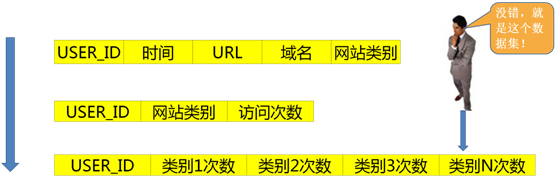
1、从19999个新用户相关的数据中选取有效数据集

2、对19999个新用户进行细分

3、对细分后的不同人群的特征进行洞察

4、有针对性的提供优惠和产品推荐

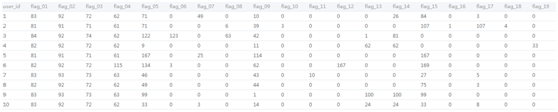
分析师经过调研，最终选取了用户的手机上网日志数据作为分析的数据集。 该数据集对用户一个月的手机上网日志数据进行清洗处理后，跟据用户ID和网站类型进行统计。 输出的是用户当月在每一类网站的访问次数。



本实验的数据集共由20列组成，一列为user\_id，剩下的19列为19个网站分类对应的一个月的访问次数，用户数19999个。网站分类情况如下：

|  |  |  |
| --- | --- | --- |
| **分类名称** | **code** | **对应字段** |
| 游戏/体育 | 1001 | flag\_01 |
| 游戏/文字 | 1002 | flag\_02 |
| 阅读/小说 | 1003 | flag\_03 |
| 阅读/漫画 | 1004 | flag\_04 |
| 网购/团购 | 1009 | flag\_05 |
| 网购/海外购物 | 1011 | flag\_06 |
| 网购/综合电商 | 1012 | flag\_07 |
| 网购/垂直电商 | 1013 | flag\_08 |
| 商旅出行/旅游咨讯 | 1026 | flag\_09 |
| 商旅出行/民宿 | 1029 | flag\_10 |
| 商旅出行/酒店门票 | 1030 | flag\_11 |
| 影音/电台 | 1035 | flag\_12 |
| 影音/音乐 | 1037 | flag\_13 |
| 影音/摄影 | 1038 | flag\_14 |
| 金融理财/金融咨讯 | 1054 | flag\_15 |
| 金融理财/P2P | 1056 | flag\_16 |
| 金融理财/证券 | 1057 | flag\_17 |
| 教育/成人教育 | 1058 | flag\_18 |
| 教育/中小学教育 | 1059 | flag\_19 |

数据集的样例数据如下：

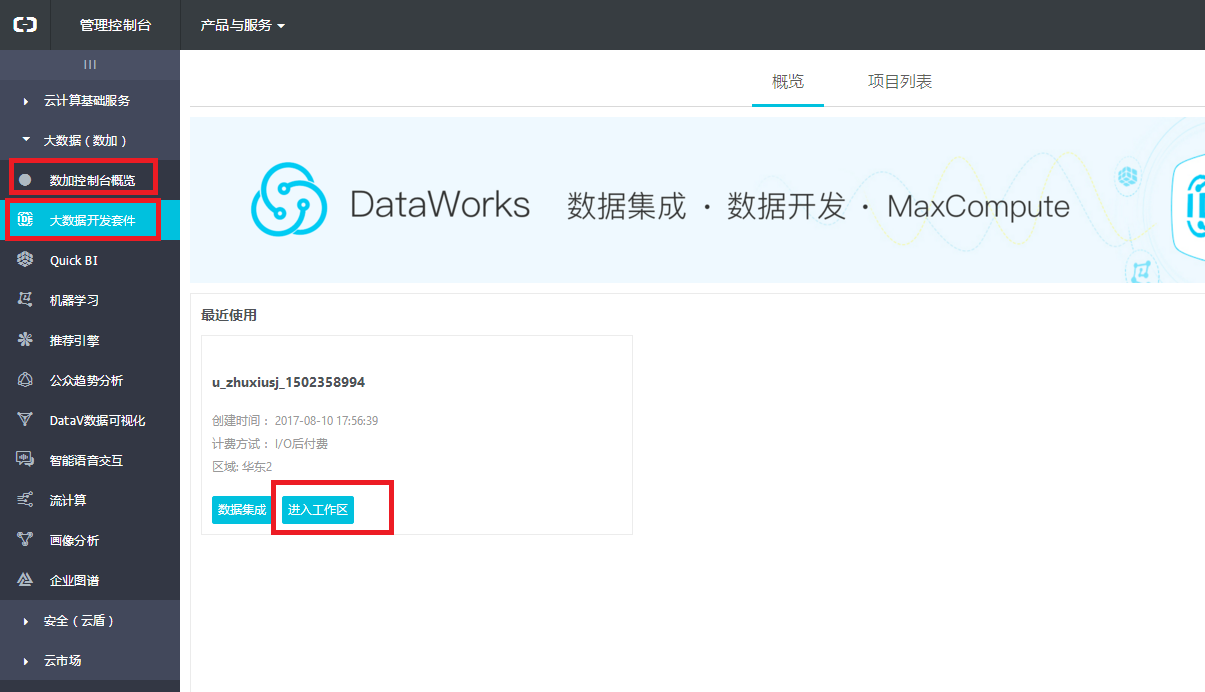


通过对该数据集进行预处理（过滤异常值、填充缺失值、归一化），将该数据集输入聚类模型，通过调整模型参数，迭代尝试不同的聚类个数，选取最优的聚类，输出结果。 对聚类结果进行业务解读，执行后续的客户经营策略。

1. 上传实验数据集

本实验中共用到1个数据集，以csv格式的静态文本提供。后续数据处理均基于这1个文件，本部分用户需通过MaxCompute数据上传的功能，将这个文件上传成MaxCompute中，作为后面试验的数据集。（也可以直接利用机器学习创建表，将数据直接上传到机器学习所建的表中）

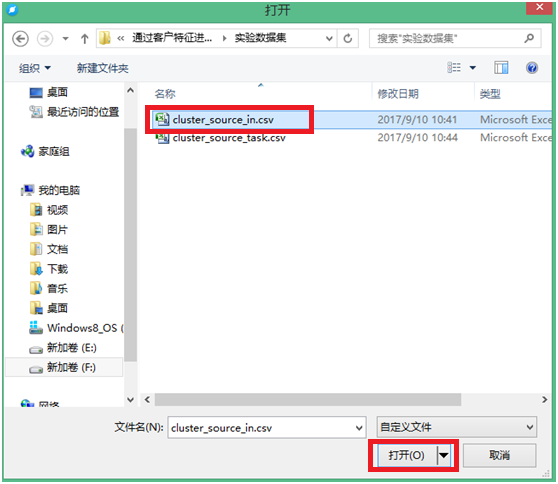
1、  点击大数据开发套件，选择刚创建的项目或已经存在项目，点击“进入工作区”；



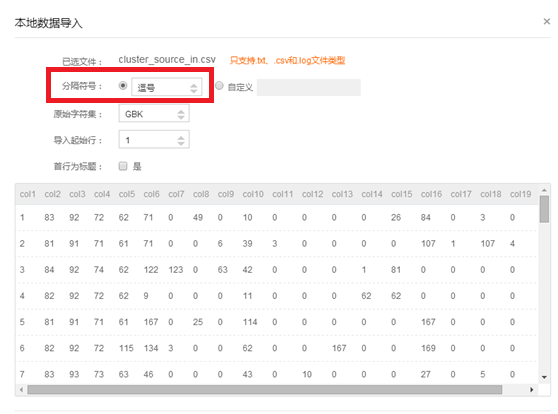
点击数据开发-> 导入-> 导入本地数据，开发区域上面菜单的“导入”，选择“导入本地数据”。



2、  点击导入的本地数据文件；



3、  打开导入的文件；选择分隔符号（默认逗号）、原始字符集、导入起始行、首行为标题 ,然后选择下一步



4、  选下一步，如果已经建表，选择已经建立的表，如果没有则新建数据表，



5、  点击新建表，输入建表语句



建表的SQL，如下：

CREATE TABLE cluster\_source\_in (

user\_id BIGINT,

flag\_01 BIGINT,

flag\_02 BIGINT,

flag\_03 BIGINT,

flag\_04 BIGINT,

flag\_05 BIGINT,

flag\_06 BIGINT,

flag\_07 BIGINT,

flag\_08 BIGINT,

flag\_09 BIGINT,

flag\_10 BIGINT,

flag\_11 BIGINT,

flag\_12 BIGINT,

flag\_13 BIGINT,

flag\_14 BIGINT,

flag\_15 BIGINT,

flag\_16 BIGINT,

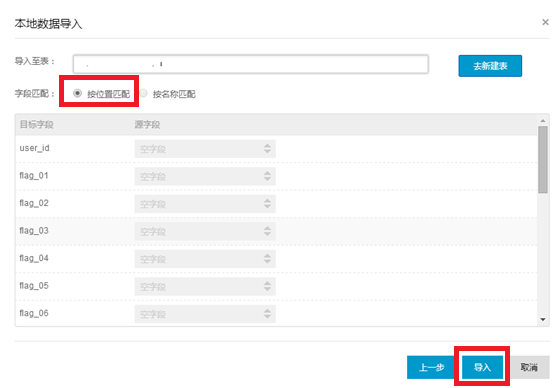
flag\_17 BIGINT,

flag\_18 BIGINT,

flag\_19 BIGINT

);

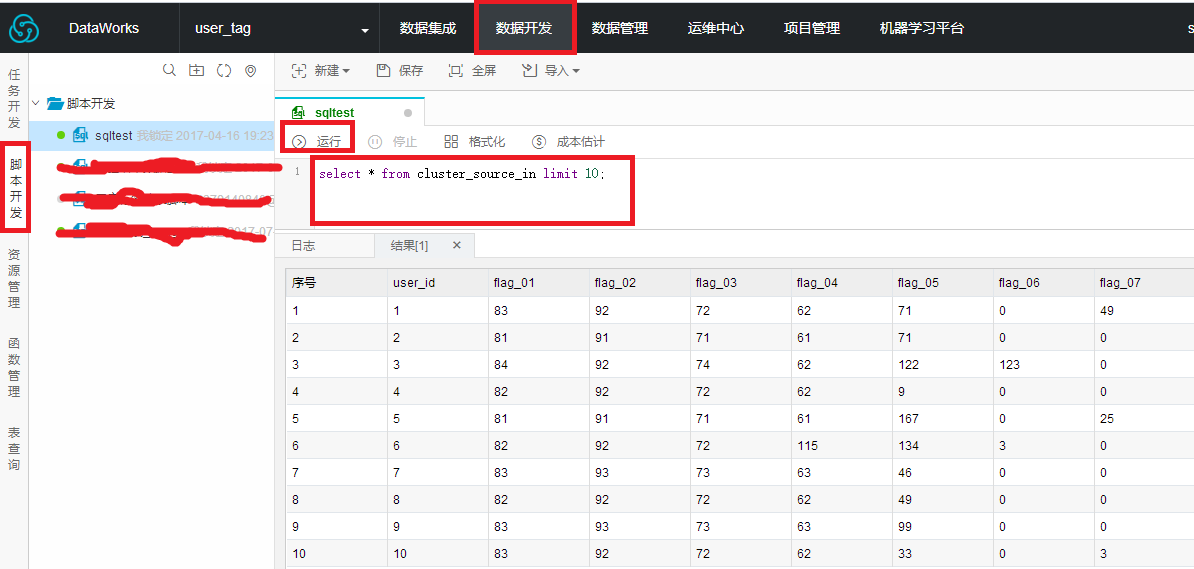
6、建表成功后，输入表名，数据映射如下



7、查看导入的数据

点击数据开发->脚本开发，新建脚本，查询新建表中的数据。

SQL脚本如下：select \* from cluster\_source\_in limit 10;



1. 创建实验

1、 在控制台左侧导航栏里点击大数据（数加），点击机器学习，进入机器学习登陆页





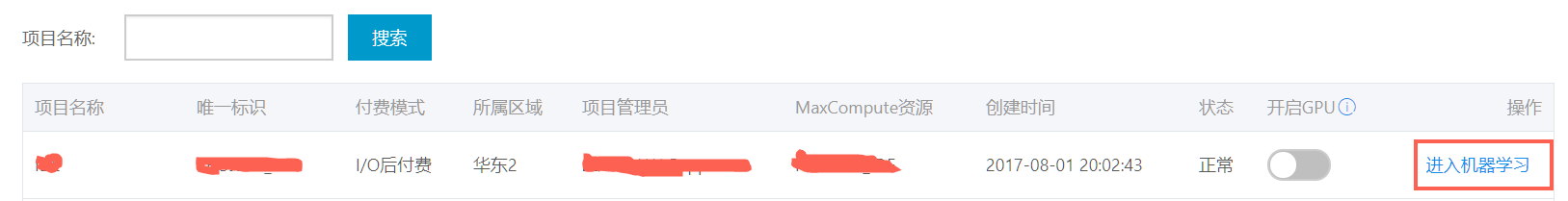
【注】如果大数据（数加）下没有项目，可以在机器学习界面创建新项目，否则可以选择一个项目作为数据实验项目；按照下列步骤新建项目：

点击控制台中的创建项目，填写相关信息，点击确认即可；如下页面：



选择付费方式（根据自己的实际情况），输入自己的项目名称、显示名以及项目描述 ，然后点击确定，创建项目。

2、 选择已创建的项目，点击进入机器学习



机器学习产品首页：



可以新建空白试验，也可以使用试验模板

3、  点击新建空白实验



给定实验名称，选择对应的项目，设定实验在我的实验中的目录位置。



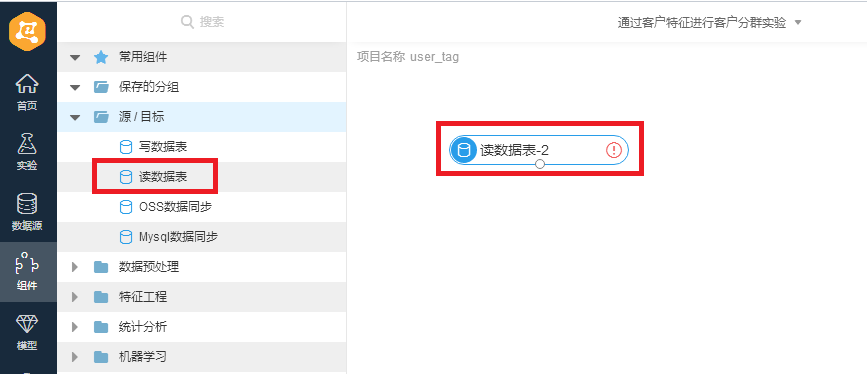
进入实验设计页面，可以开始进行实验了。

1. 数据预处理

本部分将数据导入实验组件，进行数据的预先分析、处理，目的是理解数据，了解数据。首先导入实验数据，即将数据导入到机器学习的任务组件中，然后利用PAI的数据处理组件进行数据处理

具体步骤：

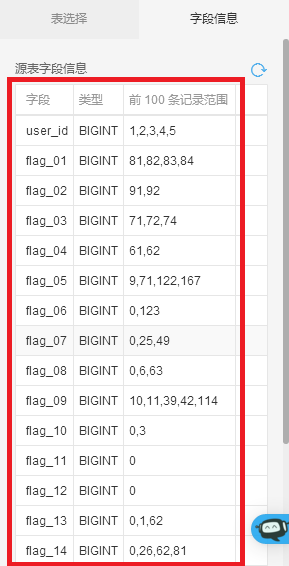
1、点击画布左侧组件栏源/目标，选择读数据组件， 拖入编辑区。



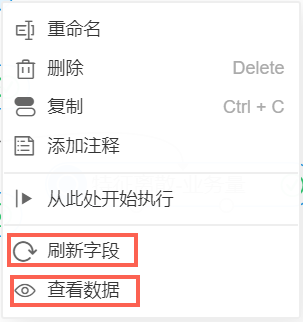
2、配置读数据表组件，双击，配置组件属性。



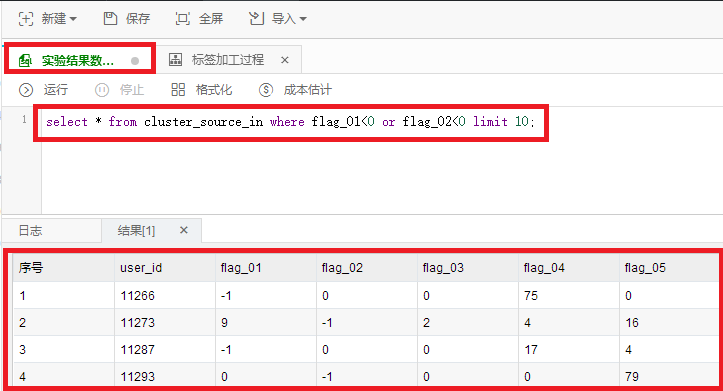
3、查看字段信息，系统自动读出数据表的字段，名称、类型、记录范围



4、右键点击读数据表组件，可以修改组件名称、刷新字段、添加注释



5、异常值过滤，通过数据探查发现flag\_01和flag\_02中有小于0的值，网站访问次数不可能小于0，因此属于异常值，本实验采用的手段是将异常值记录从数据集中去除，进行过滤操作。



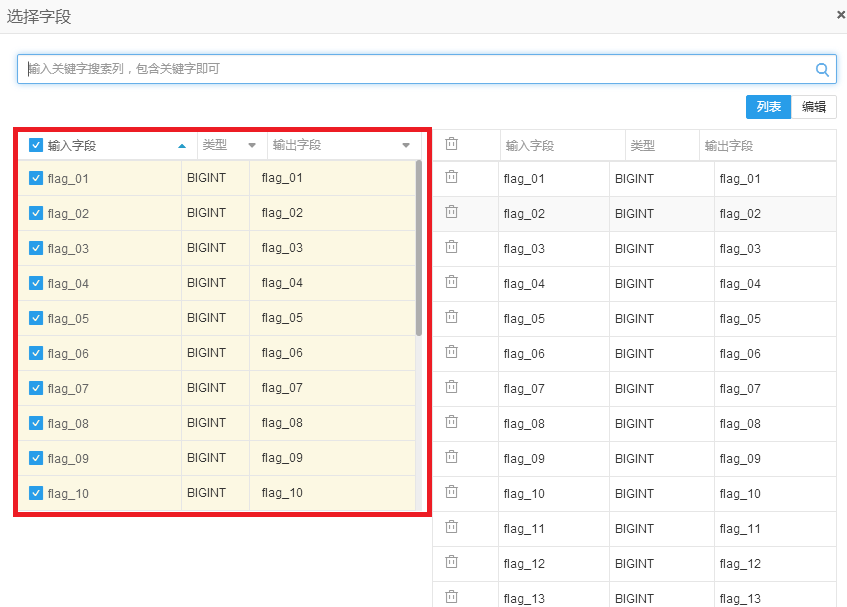
在左侧的数据预处理菜单下，选择采样与过滤/过滤与映射 组件，拖入编辑区，并将读数据表组件连线至过滤与映射组件，将读入的数据导入到过滤与映射组件进行数据过滤处理。



6、双击过滤与映射组件，配置组件属性。



点击选择字段，选取映射规则列，本实验不做列过滤，需要选择全部列。



过滤条件处填写表达式（flag\_01>=0 and flag\_02>=0）



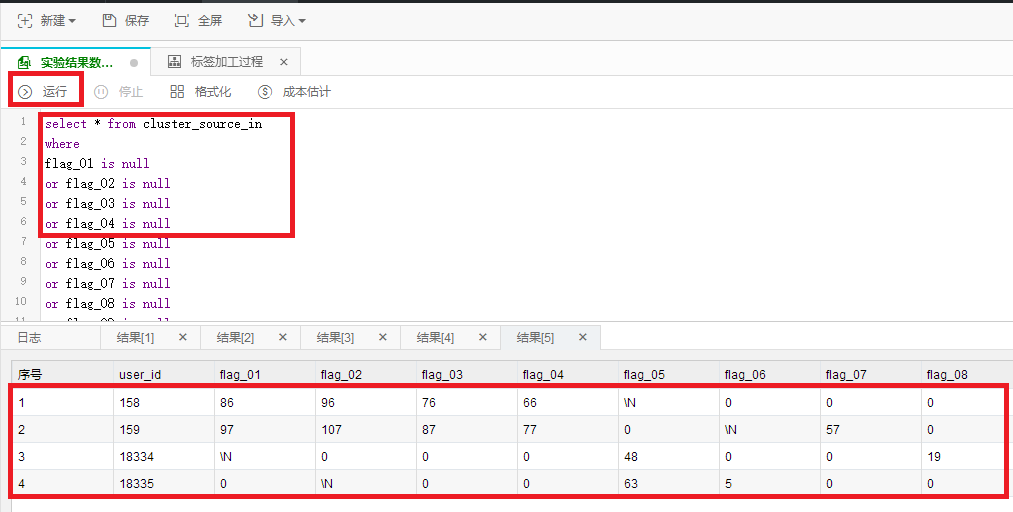
7、右点击过滤与映射组件，可以对此组件进行如下操作，执行该节点或小数据量测试该数据节点：



成功执行标志：

IMG_266

8、缺失值整理，通过数据探查发现，在flag\_01至flag\_19上，都有字段为空的可能，需要对空值进行处理。



sql检查脚本如下：

Select \* from cluster\_source\_in

Where

flag\_01 is null

or flag\_01 is null

or flag\_02 is null

or flag\_03 is null

or flag\_04 is null

or flag\_05 is null

or flag\_06 is null

or flag\_07 is null

or flag\_08 is null

or flag\_09 is null

or flag\_10 is null

or flag\_11 is null

or flag\_12 is null

or flag\_13 is null

or flag\_14 is null

or flag\_15 is null

or flag\_16 is null

or flag\_17 is null

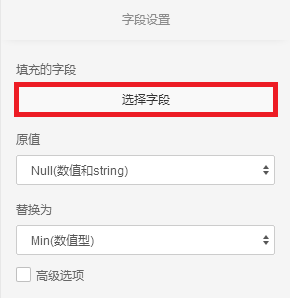
or flag\_18 is null

or flag\_19 is null

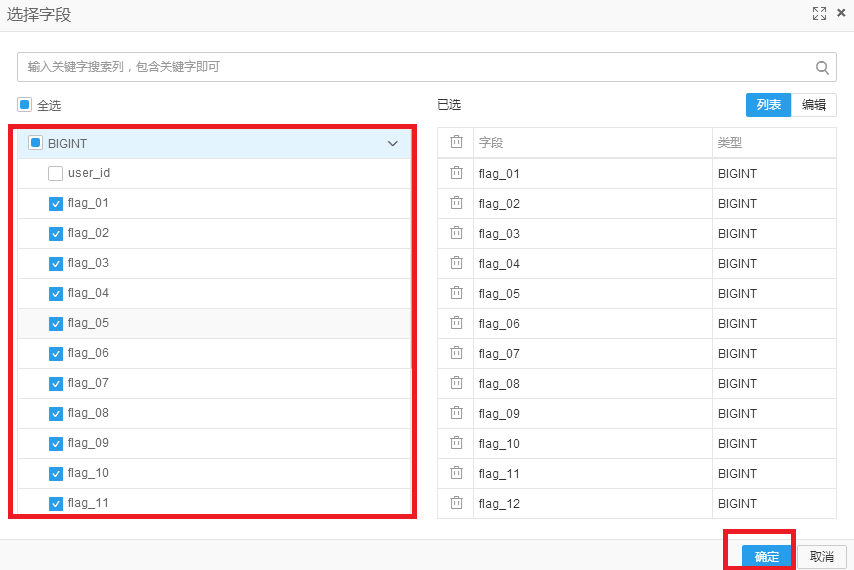
9、在左侧的数据预处理菜单下，选择缺失值填充 组件，拖入编辑区，进行数据缺失默认数值处理，从过滤与映射组件拉线至缺失值填充组件，将过滤后的数据导入缺失值填充组件进行填充（因空值在数据处理过程中会发生异常，建议优先处理）



10、双击缺失值填充组件，配置属性信息：选择要处理的字段，处理方式即为空时如何处理，是取最大、最小、平均还是自定义，类型不同处理不同。



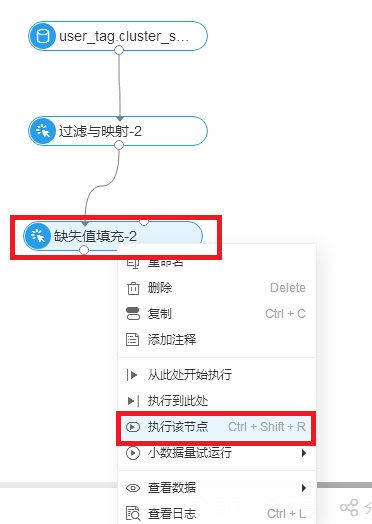
   点击选择字段按钮，选择要进行缺失值填充的字段，本实验选择19个flag字段。



设置将原值为null的替换为0。



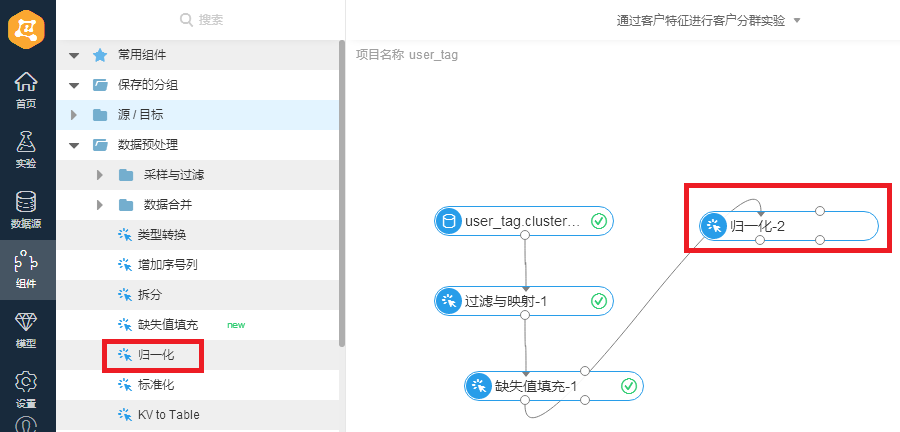
11、右点击缺失值填充组件，可以对此组件进行如下操作，执行该节点或小数据量测试该数据节点：



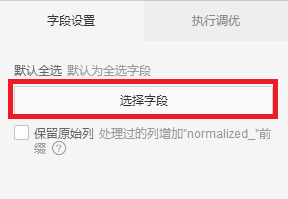
成功执行标志：

IMG_273

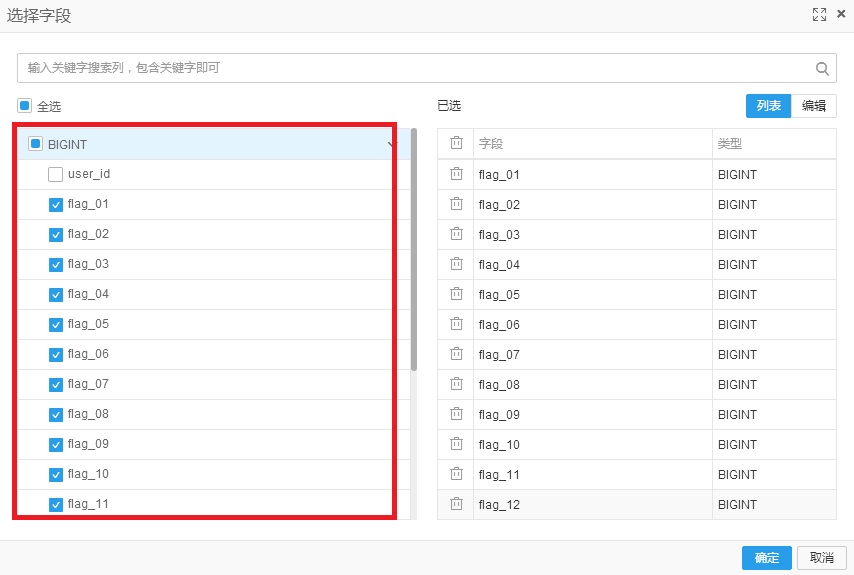
12、在左侧的数据预处理菜单下，选择归一化 组件，拖入编辑区，进行数据归一处理，将不同列的值都归一到同一量纲之下。将缺失值填充组件的输出连线至归一化组件的输入。

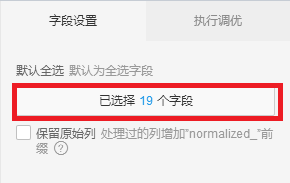


13、双击归一化组件，配置属性信息，选择要进行归一处理的字段。

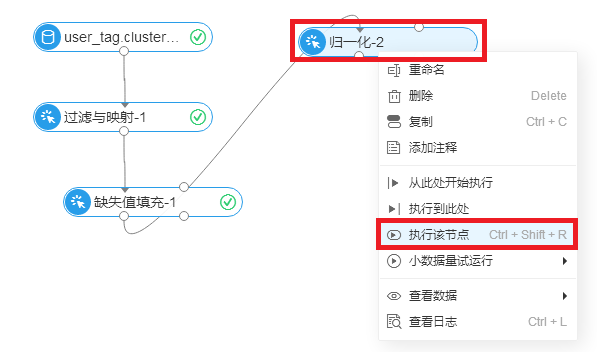


点击选择字段，选取flag\_01-flag19的字段进行归一化处理。





14、右键点击缺失值填充组件，可以对此组件进行如下操作，执行该节点或小数据量测试该数据节点。



成功执行标志：

IMG_279

1. 模型构建

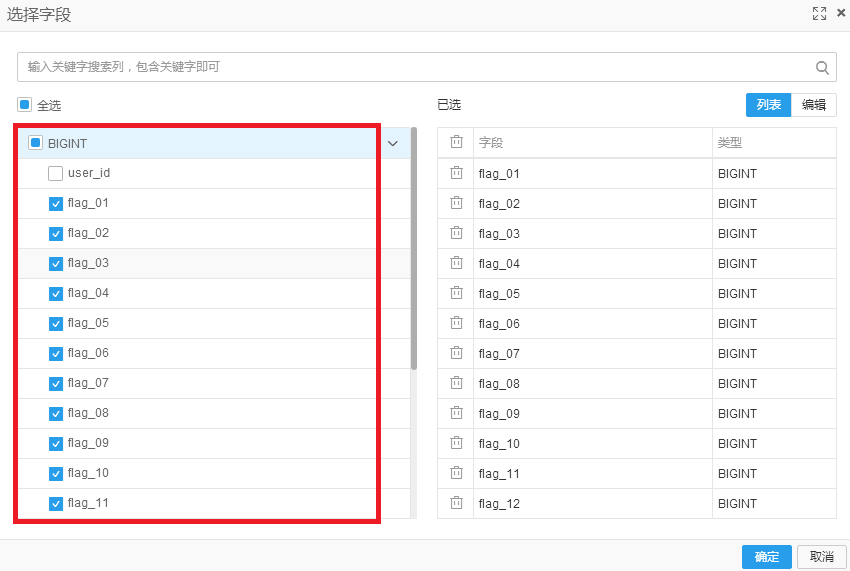
1、  选择机器学习算法，左侧菜单选择机器学习下的聚类算法，将数据处理组件K均值聚类拖入编辑区，将缺失值填充组件连线至K均值聚类组件，将填充后的数据导入K均值聚类组件。



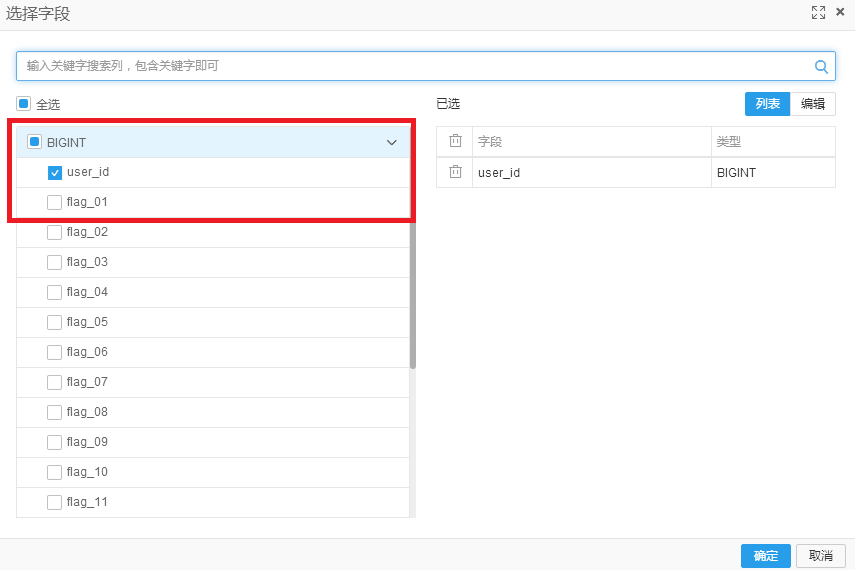
2、  双击K均值聚类组件，配置属性信息。



点击特征列->选择字段按钮，将flag\_01-flag\_19设定为特征列。



点击附加列->选择字段按钮，将user\_id设定为附加列。



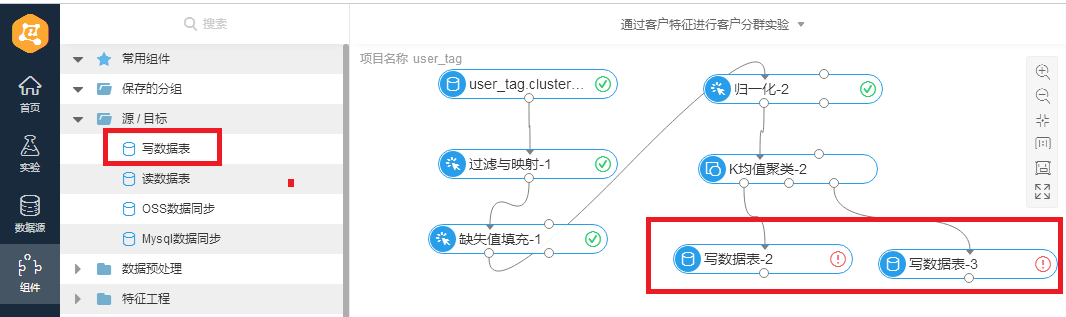
设定完毕后的界面如下：



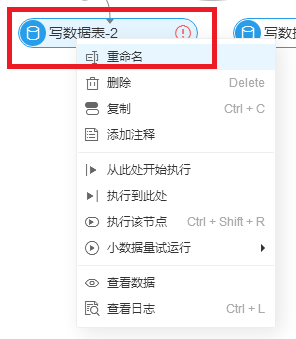
点击参数设置tab页，设置参数信息。

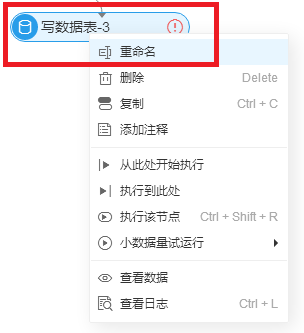


3、  分别在左侧菜单源/目标 下拖入两个写数据表组件，作为K均值聚类的输出，并连线。



左边为输出模型聚类明细表，右边为输出聚类统计表。分别右键单击两个写数据表，选择重命名，将控件重命名为聚类明细表和聚类统计表。

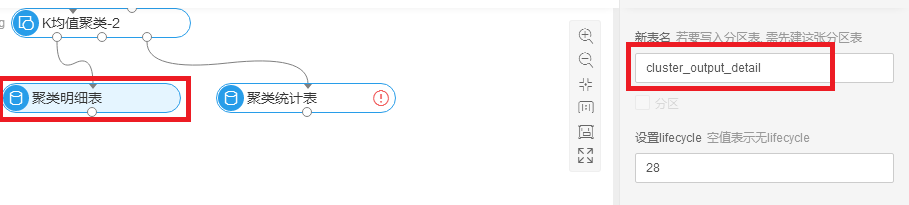


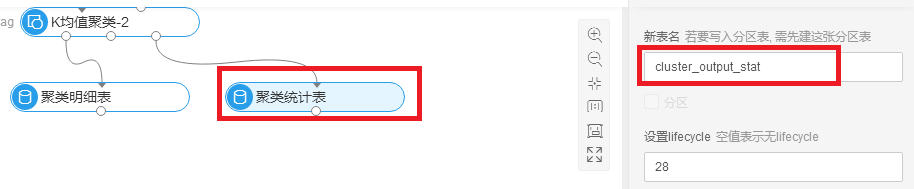




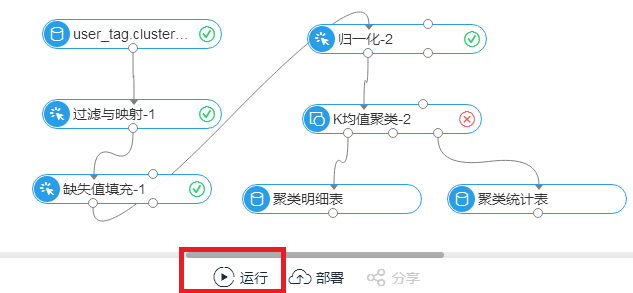
4、  配置2个写数据表组件的属性信息

双击聚类明细表控件和聚类统计表控件，分别填写它们的属性信息-表名。

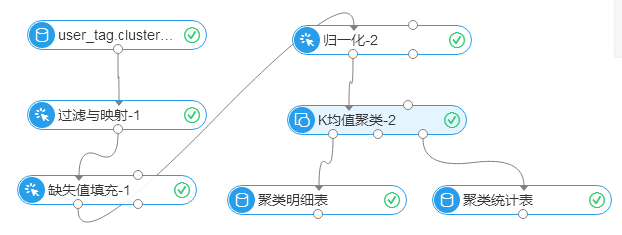




5、  点击运行按钮，执行整个处理过程



运行完毕，所有控件都打上了绿色的对勾。



6、  查看结果，通过聚类明细表和聚类统计表查看聚类结果





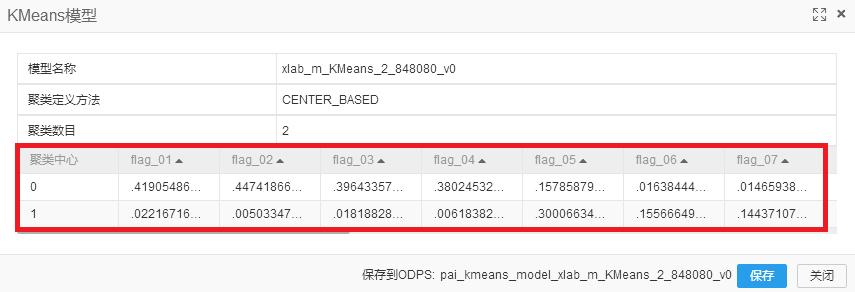
7、  查看输出的模型



点击模型描述，查看模型参数。



点击查看模型，查看模型数据



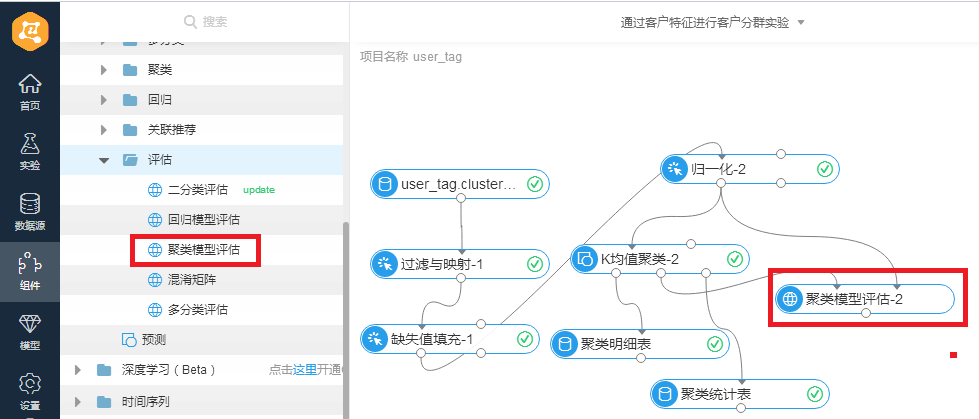
1. 模型评估

模型评估是通过输入原始数据和聚类模型，评价聚类模型的优劣，包含指标和图标

1、  在左侧菜单栏。从用机器学习进入评估，将聚类模型评估处理组件拖入编辑区。

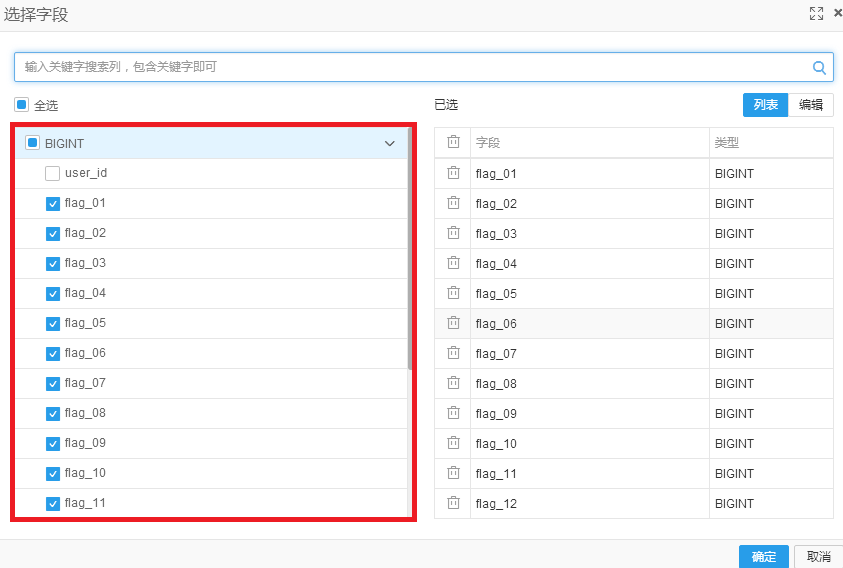


2、  将聚类模型输出以及归一化组件输出作为评估组件的输入，用连线连接起来，如图：



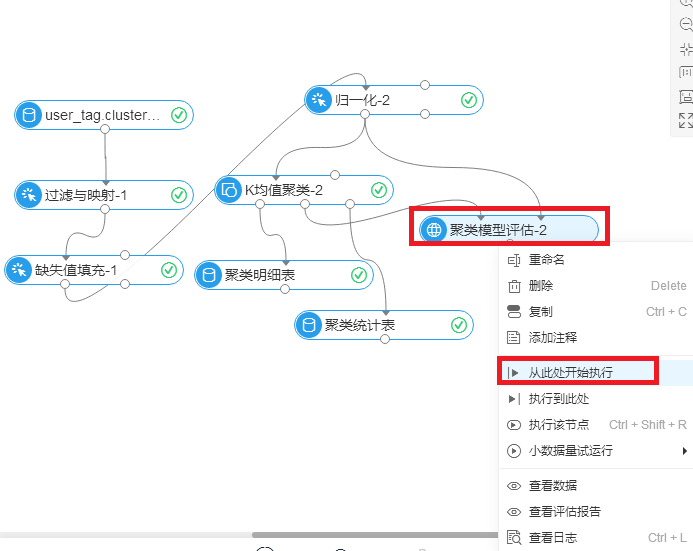
3、  配置聚类模型评估组件的字段设置，注意要和聚类组件保持一致



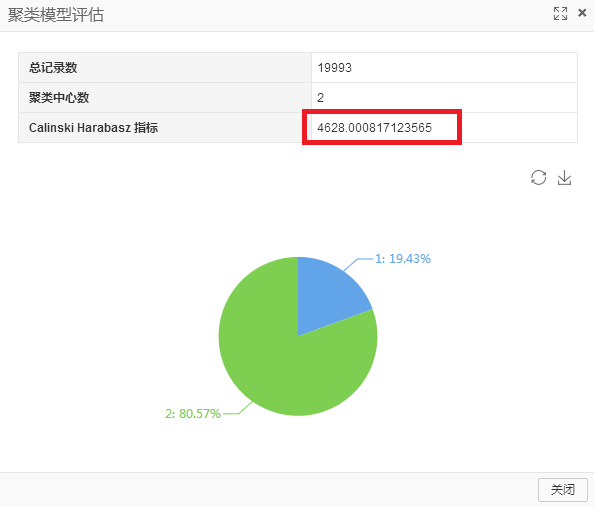
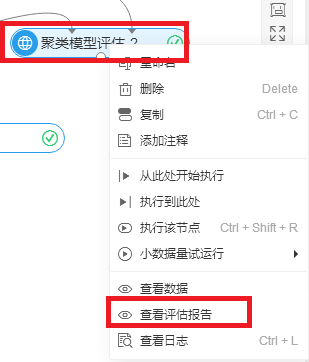




4、  右击聚类模型评估组件，选择从此处开始执行



5、  右击聚类模型评估组件，查选择看评估报告



聚类评估指标为4628，此指标越大聚类效果越好。

6、  将K均值聚类组件的聚类数参数设定为2、3、4、5、6，分别跑出最终的Calinski Harabasz 指标，选取指标值最大的类别数，作为聚类的类别个数。



|  |  |
| --- | --- |
| 聚类个数 | Calinski Harabasz 指标 |
| 2 | 4628 |
| 3 | 12156 |
| 4 | 30169 |
| 5 | 23056 |
| 6 | 22371 |

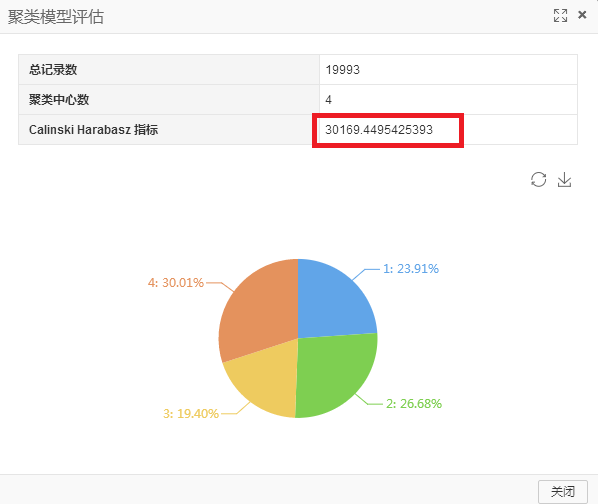
 当聚类个数为4时，Calinski Harabasz 指标值最大，最终聚类个数设定为4。

1. 结果数据查看

1、    查看聚类个数为4时，评估组件的输出

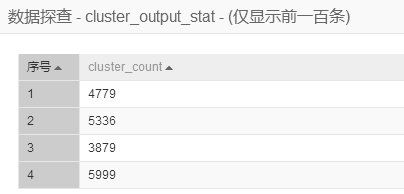


通过评估组件，可以查看四个类的占比情况



2、    查看聚类个数为4时，统计表和明细表的输出

查看统计表输出，可以获取每类的用户数

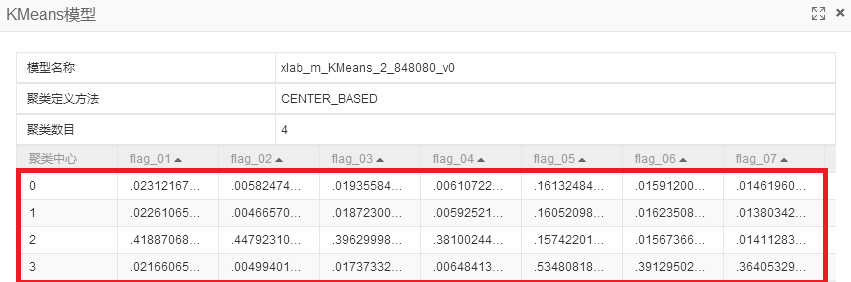


查看明细表输出，可以获取每个用户对应的类别



3、    查看聚类个数为4时，聚类模型的输出





模型给出了四个中心点的值，通过对四个中心点的值进行对比解读，提取各类的特征

1. 结果数据解读

1、    对聚类模型输出的四个聚集点中心的数据进行解读

将四个中心点的数据复制粘贴到Excel中，通过选择性粘贴->转置操作，将数据粘贴为竖列，每一列对应一个聚类，每一行对应一个网站分类，将每一行中的最大值底色标红，就可以看出每个类别的显著特征。

  
通过对上表数据进行解读，提取四类人群的特征如下：

|  |  |  |
| --- | --- | --- |
| **聚类** | **人数** | **特征** |
| **人群1** | 4779 | 该人群重视理财和教育 |
| **人群2** | 5336 | 该人群偏好旅游，喜欢音乐和摄影 |
| **人群3** | 3879 | 该人群偏好游戏和阅读 |
| **人群4** | 5999 | 该人群偏好网购 |

1. 建议方案

1、    给出建议

结合公司及其合作伙伴的业务和产品，对四类人群进行有针对性的营销推广与关怀。

针对人群1，推荐理财优惠券，推荐理财和教育专用流量包

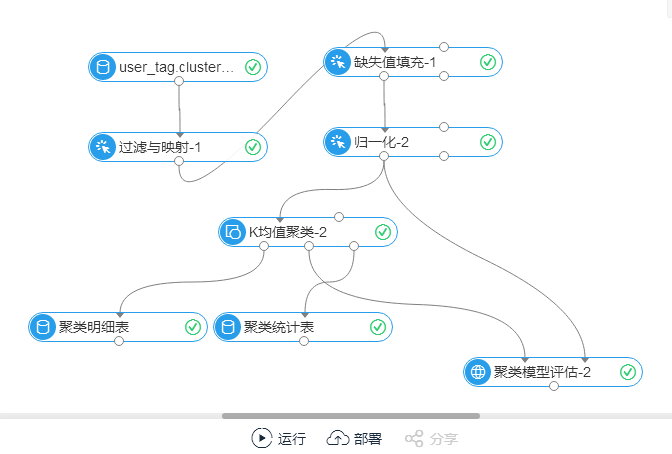
针对人群2，推荐旅游优惠券，推荐音乐视频专用流量包

针对人群3，推荐游戏卡券优惠券，推荐阅读流量包

针对人群4，推荐网购优惠券，推荐网购专用流量包

1. 实验总结

本实验完成了一个完整的数据模型构建过程，包含了业务问题解读、数据集选取、数据预处理、模型构建、模型评估、模型整体运行、部署、共享。还包括结果解读及后续的行动建议。学员可参考本实验，结合企业自身业务和需求，将所学应用到实践中去。



基本要求：学习者应该能够掌握以下知识点

1-  基本机器学习知识，如分群算法、特征处理；

2-  能根据场景和需求，结合数据特点业务特点选择合适的机器学习数据处理组件；

3-  了解机器学习的常见流程；

4-  熟悉机器学习工具PAI，能够具体数据要求进行一般的处理分析；

进阶目标：学习者通过后续的自学，可以掌握：

1-  能体会不同算法差别；

2-  可用PAI对数据进行挖掘分析，或能根据图表，进行一些辅助性的数据分析，如探索性数据分析等；

3-  了解和本行业或本公司相关的业务体系，能够结合业务知识分析数据、看懂数据；

4-  能根据自己对本行业或者本公司业务的了解，建设本行业的数据分析体系；

其他学习材料：

    阿里云官网PAI官方文档；

    机器学习相关书籍