## 注册登录御膳房

御膳房网址：<http://yushanfang.com>。请选手使用天池账号登录，登录成功后会显示如下界面：



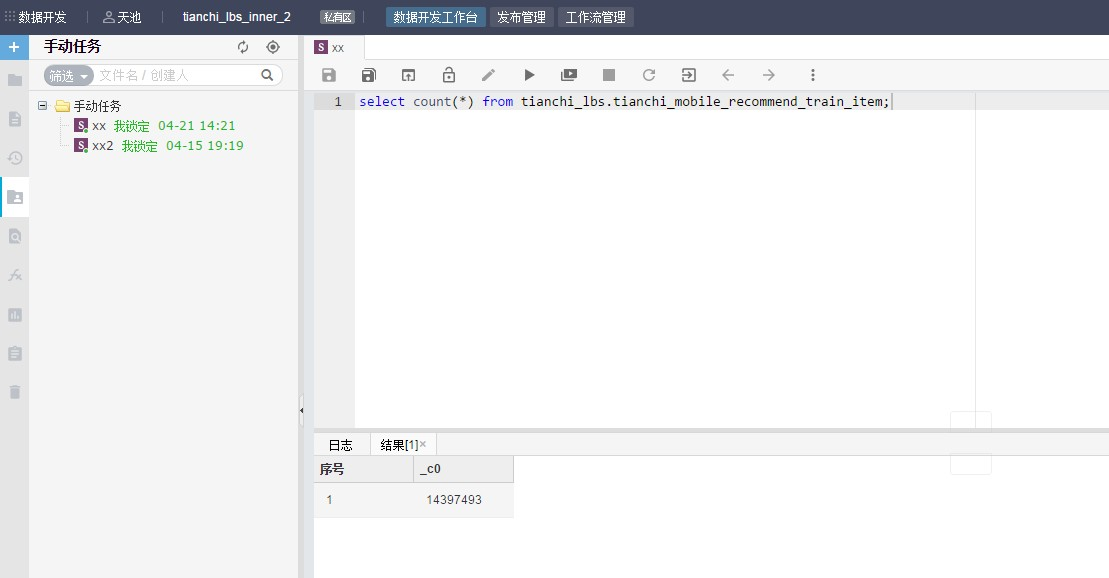
上面带红框的就是今后工作的项目空间。

## 查看浏览数据

一开始项目里面没有任何表，除了2张原始数据表。这两张表在如下的地方查看：



依次点击工作台->我的项目->数据开发，进入数据开发页面。在这里从左边工具栏可以建立各种各样的节点，也就是任务。比如我想看一下表里面的数据，建一个手动任务就可以了，如下图所示：



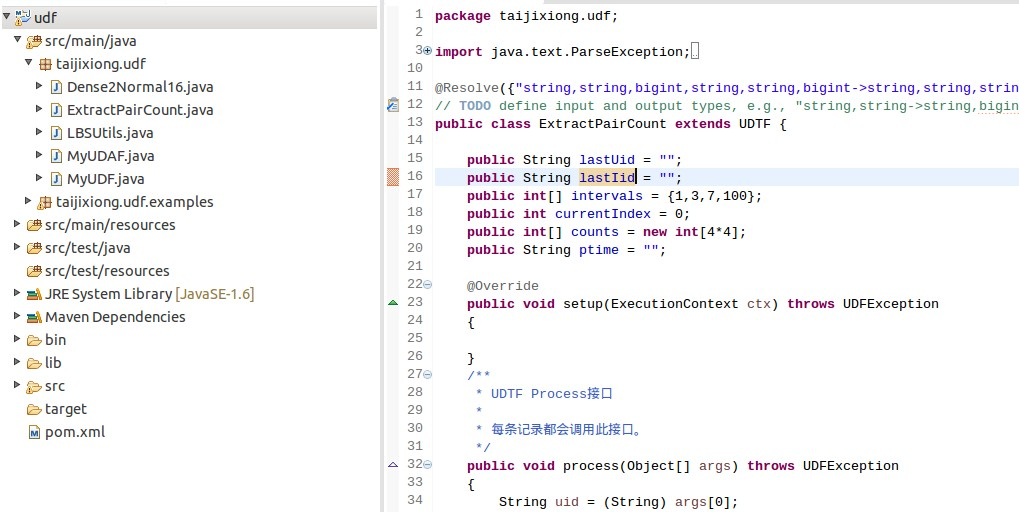
## 使用UDF

好了，经过一番数据探查，大致也确定了解题方案。假设我们采取按用户-商品对抽取特征，然后进行分类的这样一种方案，那么接下来就要抽取特征。假设我们的特征使用近1天，3天，7天，100天的用户行为数量（点击数，浏览数…），我们可以直接使用sql来抽取特征，或者使用mapreduce，再或者使用UDF，这些都可以在数据开发下添加相应的节点来完成。本文主要介绍使用UDF的方式。

UDF需要配置本地eclipse环境，文档在御膳房首页的帮助与支持下面：



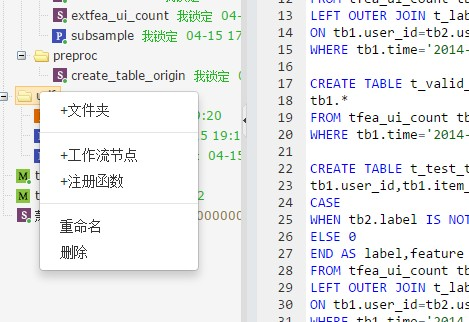
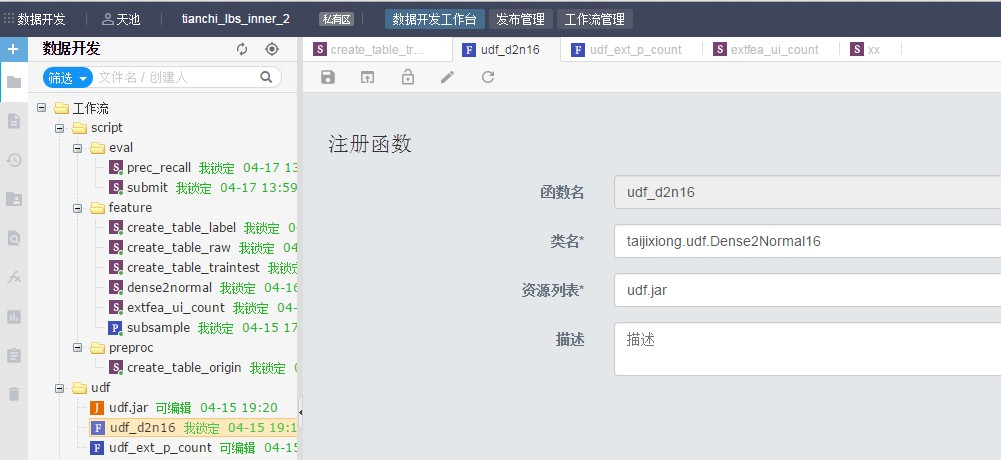
Ok，环境配好，代码写好：



按照文档所讲的方法上传到工作流下：

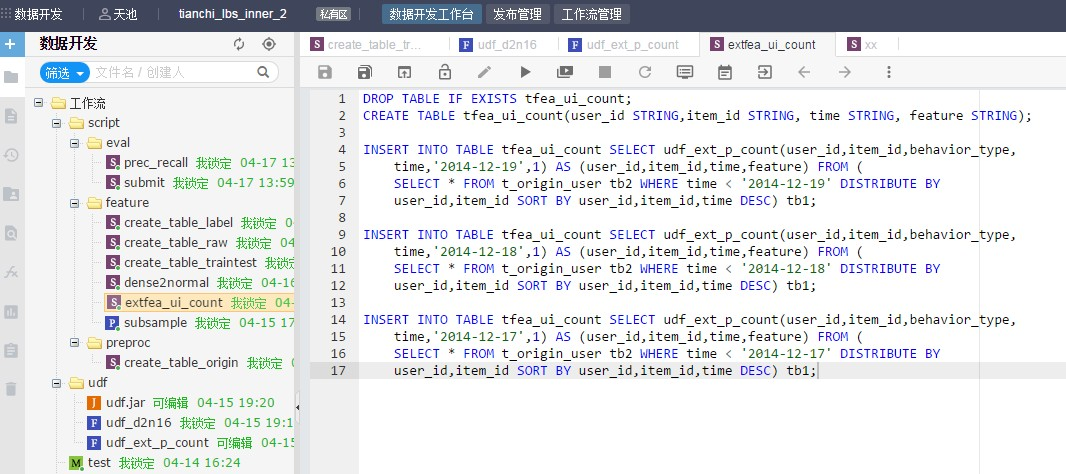


如果需要使用多个UDF，就需要手动添加注册函数：

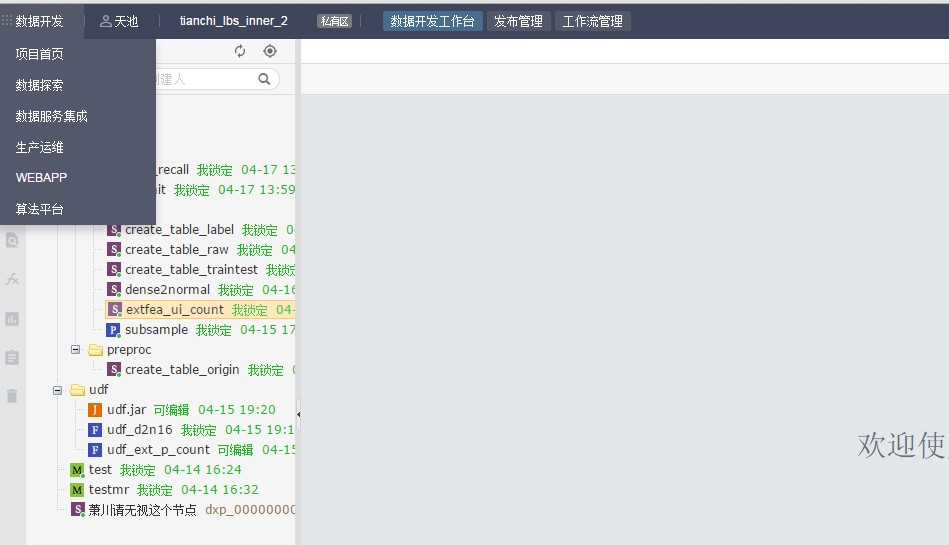
其中资源列表要填写工程生成的jar包。

下面就是用udf来抽取特征了，新建一个sql节点，代码写好后点击运行：

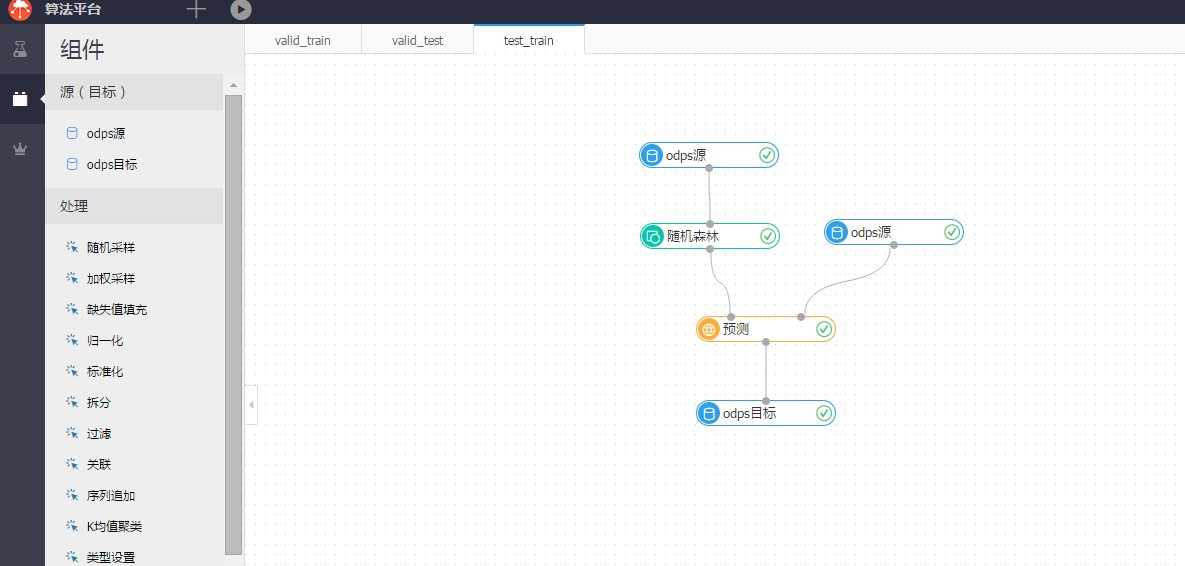


## 使用算法平台

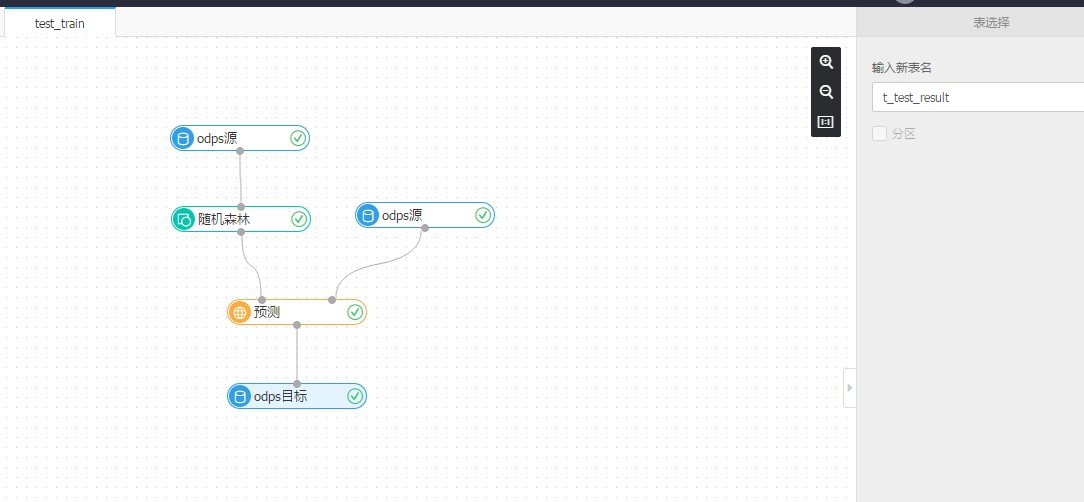
Ok，特征提好，现在就是跑模型了。跑模型除了用MR和SQL自己写以外，算法平台还提供了丰富的数据挖掘和处理算法。算法平台的入口如下：



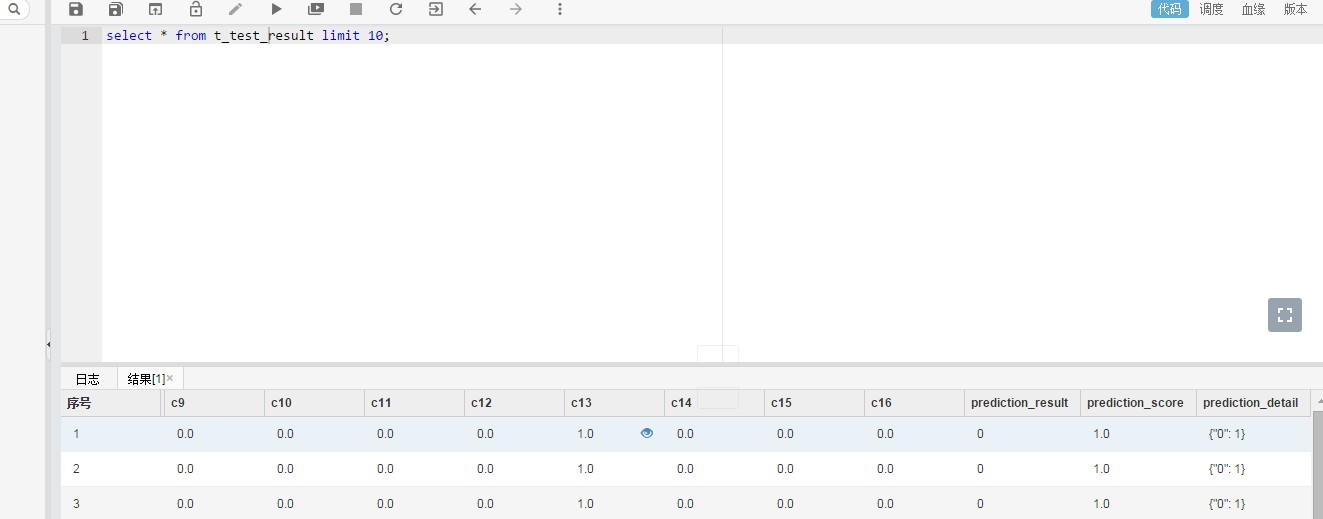
算法平台通过拖拽的方式将需要的组件拖入画布，然后用连线将各个组件的数据流链接起来，一个典型的训练预测流程如下图所示：



其中有两个odps源，分别表示训练数据和预测数据。一个odps目标表示结果表。双击组件可以对组件进行参数设置



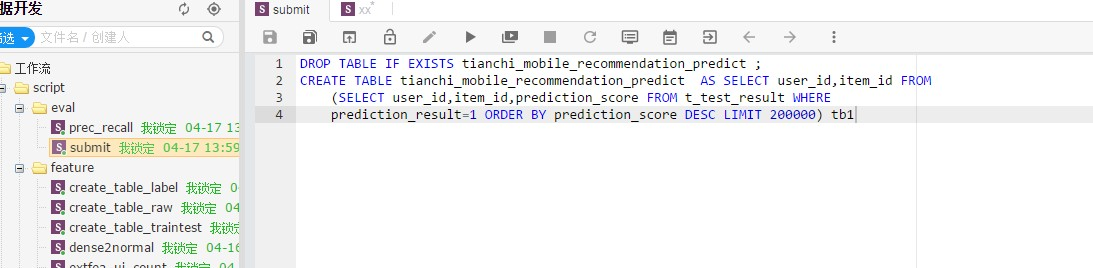
点击运行，运行完成后可以用sql查看结果表：



结果表会追加3列，predict\_result是预测结果，prediction\_score是预测为predict\_result的概率，这里要注意一下，这个概率是跟predict\_result有关的。

## 提交结果

Ok，预测结果也有了。提交只需按照概率值从大到小排序即可，如下图所示：



就这样，评测的时候评测程序会自动找到工程下的这个表名（具体请参见比赛说明），就这样我们就可以得到一个baseline的分数。祝大家玩得愉快。