# **基于天池数据集**Bank Telemarketing Campaign Case

# 探测其需要对相关客户进行营销活动的预测

**一、项目背景与挖掘目标介绍**

**1、项目背景简介**

现如今，人们想一次性购买房屋或者汽车等等价值大的东西或者是年轻人缺钱的问题等等，或多或少有需要进行信用卡贷款凑齐足够的金额，然后在慢慢的还。而由于金融机构的参与，银行可向人们发放贷款，使人们得以提前购买房屋汽车等，或是临时解决用钱量大的问题。同时，对于比如房地产开发公司来说，银行向购房者发放贷款。使购房者提前买房，有利于房地产开发公司回笼资金的周转，以获得更多的利润。对银行来说。银行为其客户提供金融服务/产品，例如储蓄账户、往来账户、借记卡等。为了增加其整体收入，该银行为其信用卡、定期存款、贷款等金融产品开展了各种营销活动。这些活动面向银行的现有客户。然而，营销活动需要具有成本效益，这样银行不仅可以增加整体收入，还可以增加总利润。

**2、项目需求与挖掘目标**

项目需求：银行为增加整体收入，为增加总利润，银行为其客户提供金融服务/产品，如贷款等金融产品。但贷款需要去拉动客户来本银行贷款，所以需要向有需求的客户介绍本公司的产品

挖掘目标：从客户的一些特征如工资，婚姻状况，是否有房屋等等相关特征，去构建模型训练，找出相关的特点。然后预测出可能需要贷款的人。再由其他人员向其介绍贷款的产品等。

**二、数据存储**

创建数据库代码:

create database yjr201806140014;

创建数据库表名代码：

create table bank\_marketing(

id string,

age string,

wage string,

balance string,

wedding string,

work string,

special string,

default\_record string,

house string,

loan string,

contact string,

last\_time string,

month string,

duration string,

number string,

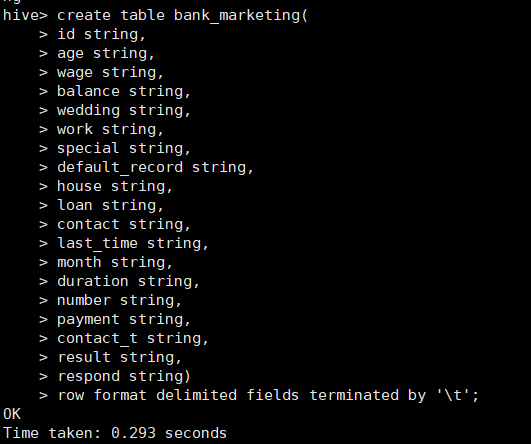
payment string,

contact\_t string,

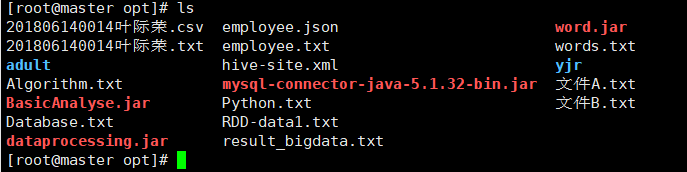
result string,

respond string)

row format delimited fields terminated by '\t';

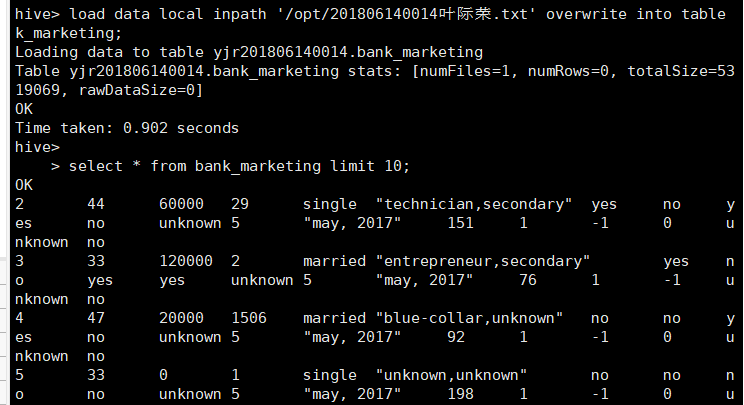


将源数据导入opt目录下

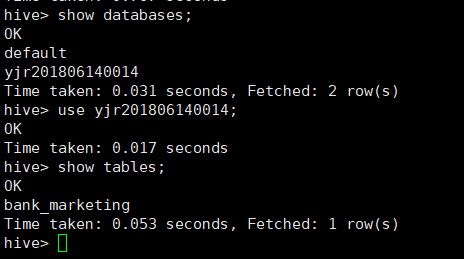


将数据导入到hive数据库名为yjr201806140014里表名为bank\_marketing中

并取出部分数据查看



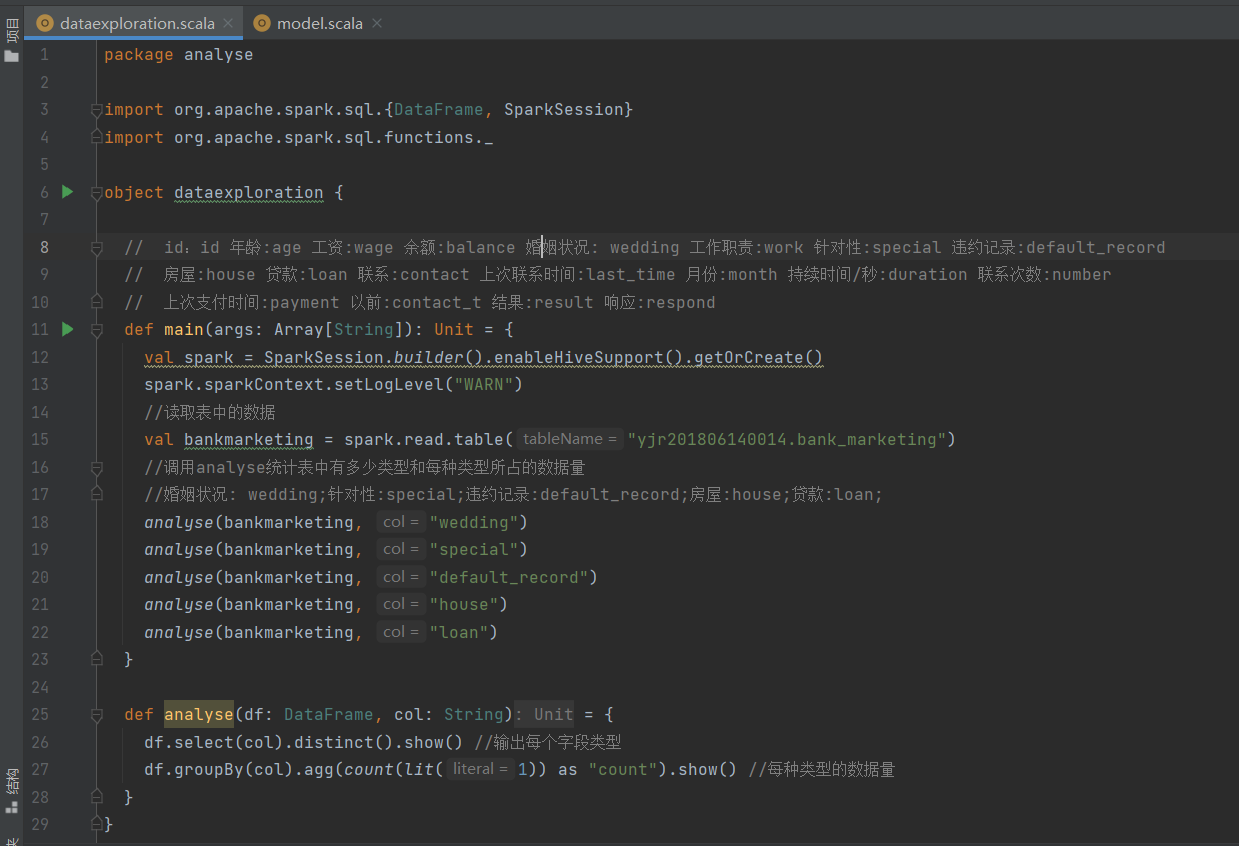
查看创建的数据库名和表名



**三、数据探索和数据预处理**

**1、数据探索源代码、注释说明及运行结果截图**

数据探索源代码截图，探索部分字段有多少种类别以及每种类别的数量

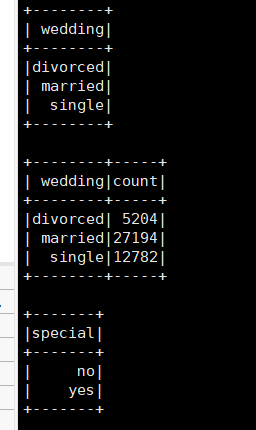


打包成jar包放在opt目录下

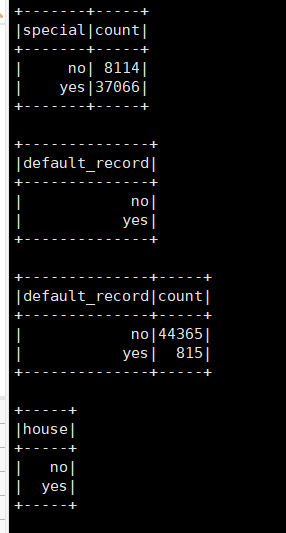
用spark--submit运行其jar包的dataexploration。



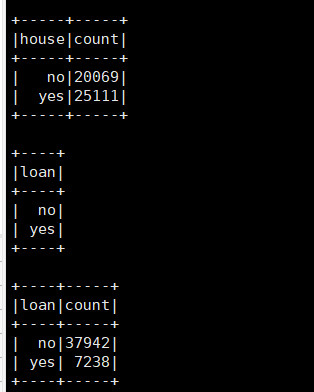
运行结果如下：wedding是婚姻状况，类别以及其数量



special是针对性，是其银行前面对客户打的标签。default\_record是违约记录。预测数据需要剔除有过违约的用户

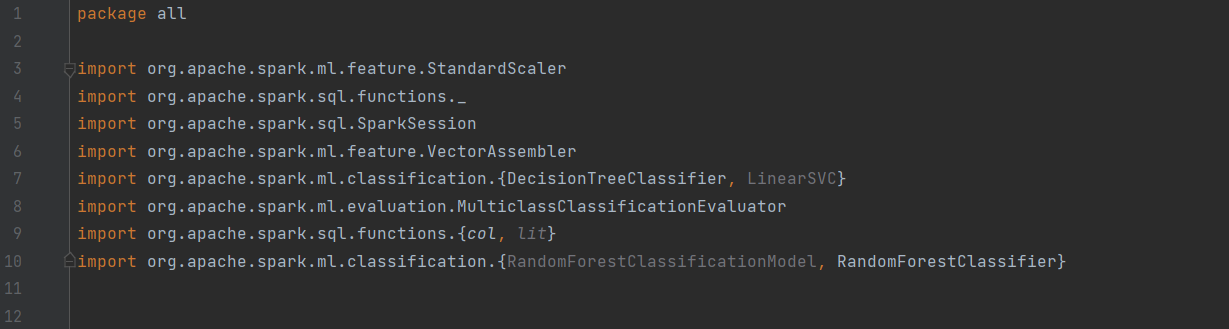


house是该客户是否有房屋。loan是该客户是否进行过贷款。可作为预测目标。



1. **数据预处理源代码、注释说明及运行结果截图**

导包



创建一个数据去存储需要构建模型的特征，id，工资，婚姻状况，针对性，违约记录，房屋，贷款。

数据探索的时候没有发现空值所以只需要剔除其有过违约记录的用户。

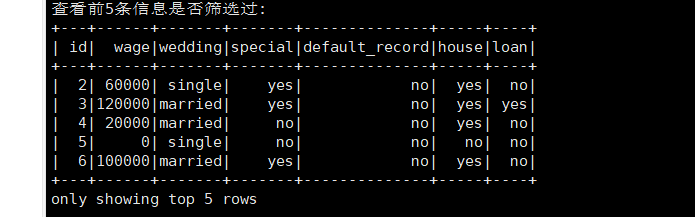


打包成jar包放在opt目录下

用spark--submit运行其jar包的model。



查看其信息是否筛选过

**四、模型构建、训练、预测、评估与应用**

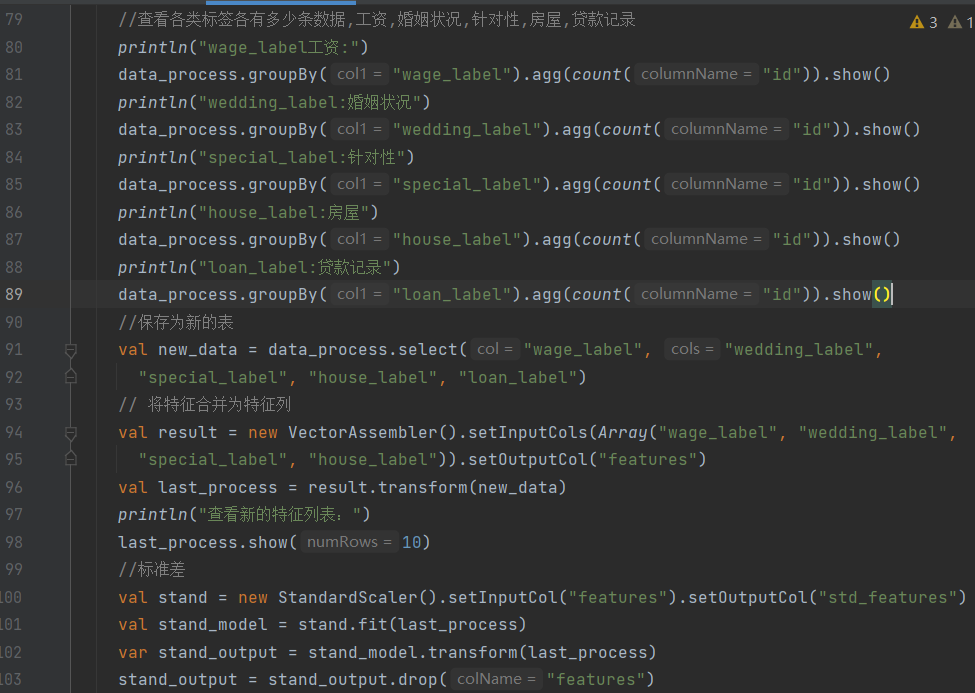
1. **数据准备、特征构建及数据集划分的源代码、注释说明及运行结果截图**

自定义udf对相关数据进行打标签和统计其数量。工资分3类，2万以下的，2万到10万的10万以上的。婚姻状况根据数据探索分为3类，已婚，未婚，离婚。针对性，是否有房屋，曾是否贷款分别各自分为两类。每种情况分类时的标签在源数据再创一列放其标签。

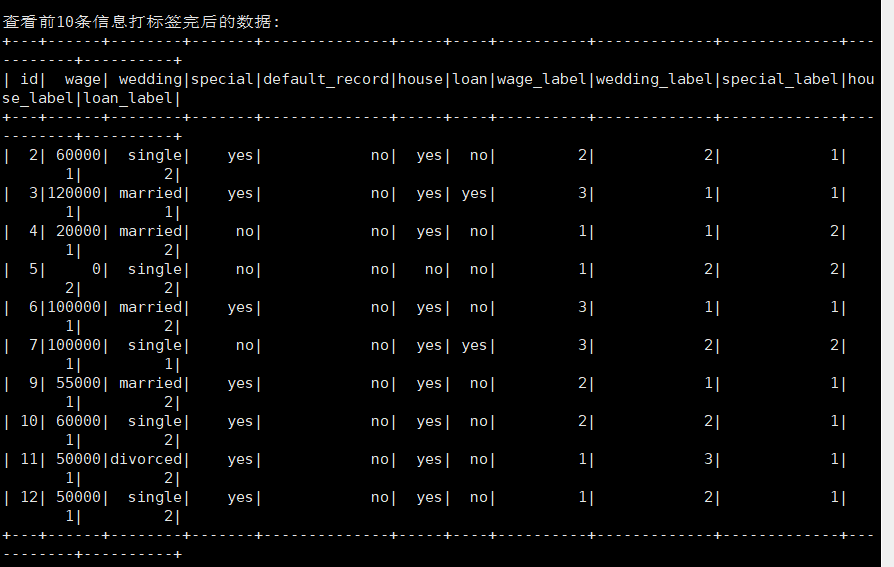




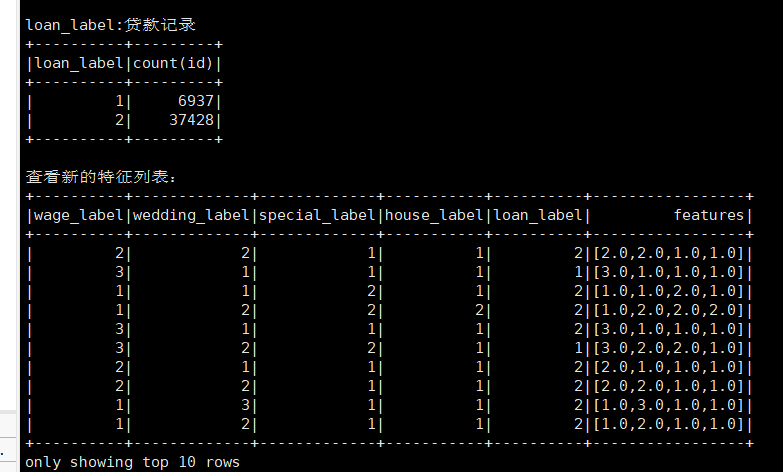
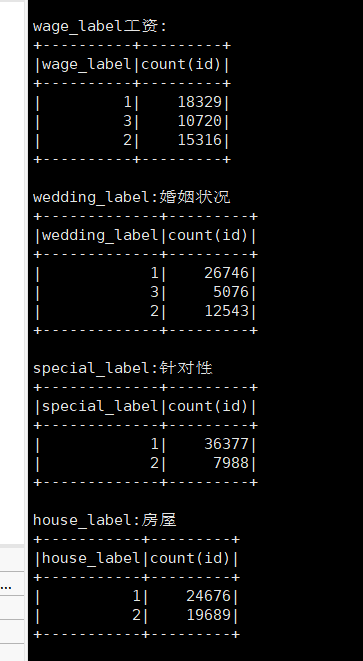
查看各类标签数据量。并创一个表放标签类。对所有的特征合并为一个特征列，然后对特征列进行标准化标准差。



对特征打完表签后的部分数据：

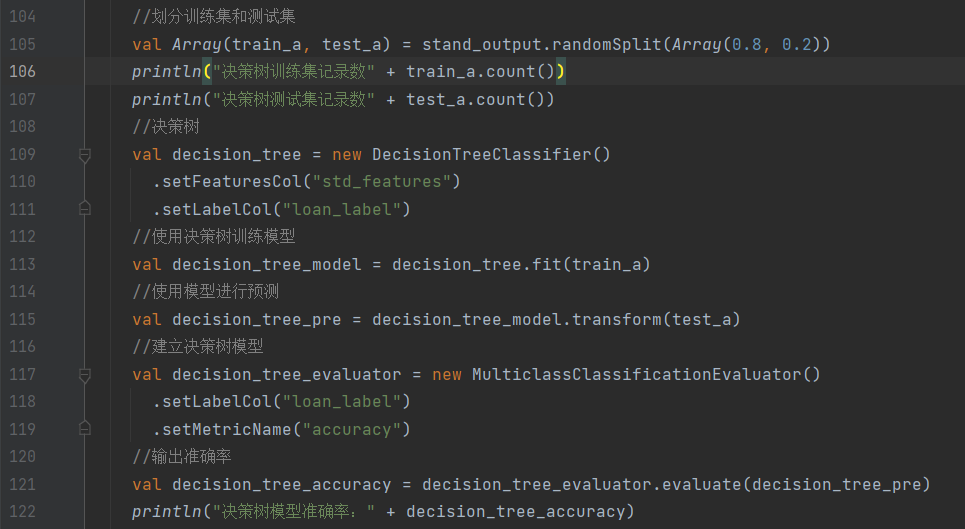


查看各类标签数据量。以及特征合并后的部分列表。

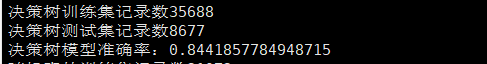


**2、第1种算法模型的构建、训练、预测、评估与推广应用的源代码、注释说明及运行结果**

第一种算法我用的是决策树



训练集和测试集的记录数以及模型预测的准确率

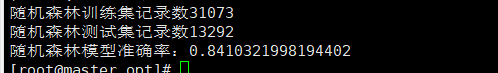


**3、第2种算法模型的构建、训练、预测、评估与推广应用的源代码、注释说明及运行结果**

第二种算法我用的是随机森林算法



训练集和测试集的记录数以及模型预测的准确率

**4、两种模型分析结果的对比与说明**

决策数应用的训练集和测试集比例为2：8；准确率为0.844

随机森林应用的训练集和测试集比例为3：7；准确率为0.841

可以看到两种模型训练测试集的比例不一样，但预测结果很像；其实是因为两种模型很像，但又不同；

决策树是一种有监督的机器学习算法，该方法可以用于解决分类和回归问题。决策树可以简单地理解为达到某一特定结果的一系列决策，随机森林是基于树的机器学习算法，该算法利用了多棵决策树的力量来进行决策。它是随机创造的决策树组成的森林。简单来说：随机森林算法用多棵（随机生成的）决策树来生成最后的输出结果。