



机器学习服务

最佳实践

文档版本 01

发布日期 2018-01-03

华为技术有限公司



版权所有 © 华为技术有限公司 2018。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址：深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址：<http://www.huawei.com>

客户服务邮箱：support@huawei.com

客户服务电话：4008302118

目录

1 银行定期存款业务预测	1
1.1 业务场景介绍	1
1.2 操作流程概述	3
1.3 业务数据上传	3
1.4 搭建模型训练与评估 workflow	4
1.5 训练、评估和筛选模型	8
1.6 保存模型	11
1.7 搭建模型预测 workflow	12
1.8 预测数据	13
2 商品批发商客户分群	15
2.1 业务场景介绍	15
2.2 操作流程概述	16
2.3 业务数据上传	17
2.4 搭建模型训练 workflow	18
2.5 训练与评估模型	21
2.6 获取客户分群结果	24
3 电影推荐	28
3.1 业务场景介绍	28
3.2 操作流程概述	30
3.3 业务数据上传	30
3.4 搭建模型训练 workflow	31
3.5 训练模型	33
3.6 搭建模型预测 workflow	33
3.7 获取评分结果	35
A 附录	37
A.1 修订记录	37

1 银行定期存款业务预测

使用 workflows 进行模型训练并保存模型文件，再使用 workflows 实现银行定期存款业务的预测。

1.1 业务场景介绍

场景描述

业务预测是企业面临的一个常见问题，主要是通过历史业务数据或者客户特征数据，对未来一定时间内的业务发展和客户办理业务意愿进行预测。

本次业务场景为：在银行行业中，银行业务人员手头上有大量的老客户的个人信息。那么，业务人员如何利用这些老客户的数据信息，根据新客户提供的个人信息，预测新客户是否愿意办理定期存款业务。然后，业务人员根据预测结果，可以为有存款意愿的客户推送相关的业务宣传材料，同时减少对无存款意愿的客户推送活动，避免引起客户的反感。

数据说明

本次业务场景的模拟数据来自UCI的Machine Learning Repository，地址为<https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>。在数据集目录下，单击下载“bank-full.csv”数据集，该数据集共有45212个样本。我们保留部分特征列（age、job、marital、education、housing、loan）和最后1列。其中：前6列数据分别代表：年龄、工作类型、婚姻状况、文化程度、是否有房贷和是否有个人贷款。最后1列表示该客户是否愿意存款，作为样本的类别标签。部分数据样例如表1 部分样例数据所示。

表 1-1 部分样例数据

age	job	marital	education	housing	loan	y
50	admin.	married	primary	no	yes	no
57	admin.	divorced	primary	no	no	no
55	admin.	married	primary	yes	no	no

age	job	marital	education	housing	loan	y
38	admin.	married	primary	yes	no	no
35	admin.	married	primary	yes	no	no
48	admin.	married	primary	no	no	no

针对该业务场景描述，我们对数据集做如下处理：

- 由于原始数据集中正负样本比例差异过大，会降低模型性能。因此，我们手动从所有样本中选取12290个样本（正样本：负样本=5290：7000，尽量保持正负样本比例在1：1左右）模拟业务人员手中已有的老客户数据，并保存为：“BANK_DATA.csv”。
- 使用表1-2的3个样本模拟银行业务人员获取3位新客户个人信息数据，并保存为：“NEW_DATA.csv”。

表 1-2 “NEW_DATA.csv” 数据集

age	job	marital	education	housing	loan
41	blue-collar	married	secondary	yes	no
53	blue-collar	married	secondary	yes	no
38	management	married	tertiary	no	yes

那么，该业务场景的问题可理解为：

银行业务人员手头上有一份老客户的数据“BANK_DATA.csv”，他如何使用机器学习服务，根据3位新客户的数据“NEW_DATA.csv”，预测这3位新客户是否具有存款意愿。

解决方案分析

通过问题描述可知客户办理定期存款只有“是”、“否”两种可能，因此该问题为分类问题中的二分类问题。分类问题可以用“朴素贝叶斯”、“逻辑回归”或“随机森林”算法来分析处理：

- 朴素贝叶斯

朴素贝叶斯算法是基于贝叶斯定理与特征条件独立假设的分类方法。对于给定的训练数据集，首先基于特征条件独立假设学习输入/输出的联合概率分布；然后基于此模型，对给定的输入x，利用贝叶斯定理求出后验概率最大的输出y。朴素贝叶斯法实现简单，学习与预测的效率都很高，是一种常用的方法。

贝叶斯分类器的分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。该模型所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。朴素贝叶斯模型假设属性之间相互独立，在属性相关性较小时，该模型具有良好的性能。

- 逻辑回归（Logistic Regression）

逻辑回归算法是一种常用的二分类算法。它的一般处理步骤如下：

- a. 将“样本属于此类”与“样本不属于此类”的概率之比转化成线性函数。
- b. 借助极大似然估计对此函数进行参数估计，求得似然函数。
- c. 利用梯度下降算法或牛顿法对似然函数最大化求解，不断迭代获得原始问题的最优解，得到逻辑回归模型。

- 随机森林（Random Forest）

随机森林算法是数据挖掘中的一种分类方法，用于产生分类模型或回归模型。它用随机的方式建立一个森林，森林由很多决策树组成，每一棵决策树之间没有关联。得到森林之后，当有一个新的样本输入时，就让森林中的每一棵决策树分别进行判断，看这个样本对应哪一类（分类）或哪一个值（回归）。对于分类问题，哪一类被选择最多，就预测这个样本为那一类；对于回归问题，取所有树的预测值的平均值。

接下来，我们将训练得到朴素贝叶斯、逻辑回归和随机森林3个模型，通过比较这3个模型的性能指标，筛选出最优模型，对新客户存款意愿进行预测。

1.2 操作流程概述

流程介绍

使用机器学习服务的工作流进行数据分类的操作过程主要包括6个步骤：

1. **业务数据上传**。将业务数据上传通过OBS和MRS，上传到HDFS中，供用户分享使用。
2. **搭建模型训练工作流**。业务人员首先根据手中的老用户数据集“BANK_DATA.csv”，训练3种算法模型，此步骤主要是完成模型训练的工作流搭建操作。
3. **训练、评估和筛选模型**。运行工作流后得到3种模型的评估结果，业务人员通过分析3种模型的评估结果，筛选出最优的模型。
4. **保存模型**。把已训练好的并具有良好的性能的模型保存到HDFS中，用于后续使用。
5. **搭建模型预测工作流**。使用得到的模型根据新客户的数据“NEW_DATA.csv”，对客户存款意愿进行预测，此步骤主要是完成模型预测的工作流搭建操作。
6. **预测数据**。运行工作流，得到预测结果并对预测结果做简要说明。

1.3 业务数据上传

使用机器学习服务前，需要将本地数据文件上传至OBS，再通过MRS将数据从OBS中导入至HDFS，供MLS从HDFS中读取数据。

上传数据至 OBS


步骤1 登录华为云管理控制台。

步骤2 单击“服务列表”，选择“存储 > 对象存储服务”。

步骤3 单击右上角的“创建桶”，进入“创建桶”页面。

表 1-3 “创建桶” 样例

参数名	样例值
区域	选择新建桶所在的区域。
桶名称	obs-mls

- 步骤4** 单击“创建桶”，创建桶成功，返回“对象存储服务”页面。
- 步骤5** 在桶列表中，选择并单击“obs-mls”，进入“桶：obs-mls”界面。
- 步骤6** 在左侧导航栏中选择“对象”，单击“上传”，弹出“上传对象”页面。
- 步骤7** 单击，在弹出框中选择待上传的数据文件“BANK_DATA.csv”和“NEW_DATA.csv”。
- 步骤8** 单击“确定”，等待页面提示上传成功。

----结束

通过 MRS 将数据导入 HDFS

- 步骤1** 单击“服务列表”，选择“EI 企业智能>MapReduce服务”。
- 步骤2** 在左侧导航栏中，选择“集群列表>现有集群”。
- 步骤3** 选择当前MLS实例所关联的MRS集群的名称，例如：“mrs-mls”。选择“文件管理”分页，单击“导入数据”。
- 步骤4** 在弹出框中，选择表1-4中的路径。

表 1-4 导入路径

路径	样例值
OBS路径	“s3n://obs-mls/BANK_DATA.csv”和“s3n://obs-mls/NEW_DATA.csv”。
HDFS路径	“/user/omm/mls”。

- 步骤5** 单击“确定”，等待页面提示导入成功。

----结束

1.4 搭建模型训练与评估 workflow

模型训练和评估 workflow 主要包括4部分内容：

- 1. 读取业务数据集。
- 2. 设置数据集中的类别标签列、并将数据拆分为训练集和验证集。
- 3. 使用训练集分别训练朴素贝叶斯、逻辑回归和随机森林三种模型。

4. 分别计算这3种模型的AUC、召回率和F1 score值等作为性能评估指标。

操作步骤

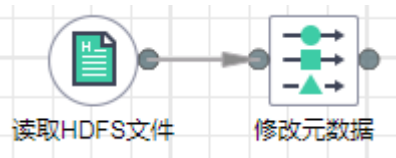
- 步骤1 新建并打开一个工作流，命名为“bank_training”，进入工作流编排界面。
- 步骤2 读取数据。使用“读取HDFS文件”节点从HDFS中读取业务数据，操作步骤如下：
将“输入”展开目录中的“读取HDFS文件”节点拖拽至画布中，单击该节点，在右侧参数配置区域按照表1-5配置参数。

表 1-5 “读取 HDFS 文件” 节点参数配置样例

参数名称	样例值
数据格式	CSV
数据文件	“/user/omm/mls/BANK_DATA.csv”
导入元数据	不勾选
是否包含表头	勾选
字段分隔符	,
保存元数据文件	不勾选
处理异常值	null替代值
保存异常记录	不勾选

- 步骤3 设置样本的类别列。使用“修改元数据”节点将“y”列设置为样本的类别标签列。该列作为样本的真实类别值，通过与后续验证集在模型中的预测结果进行比较，得到模型的评估结果。操作如下：
1. 将“数据转换 > 字段操作”展开目录中的“修改元数据”节点拖拽至画布中。如图1-1将“修改元数据”节点连接到“追加”节点。

图 1-1 连接“修改元数据”节点



2. 单击该节点，在右侧参数配置区域单击，新增一组字段，参数配置如表1-6所示。

表 1-6 “修改元数据” 节点参数配置样例

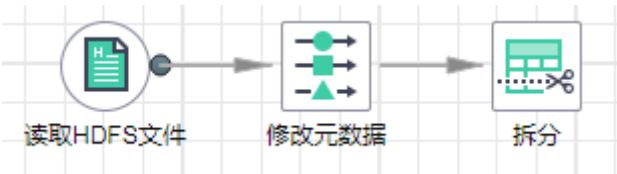
参数名称	样例值
字段	“y”
角色	“Target”

参数名称	样例值
测量尺度	“Flag”
值	“yes,no”
输入数据模式	-
输出数据模式	-

步骤4 划分训练集和验证集。使用“拆分”节点将数据集集才分为训练集和验证集。其中，66%作为训练集，剩余34%作为验证集。操作如下：

1.
- 将“数据转换 > 记录操作”展开目录中的“拆分”节点拖拽到画布中。将“修改元数据”节点连接到“拆分”节点，如图1-2。

图 1-2 连接“拆分”节点



2.
- 单击该节点，在右侧参数配置区域按照表1-7配置参数。

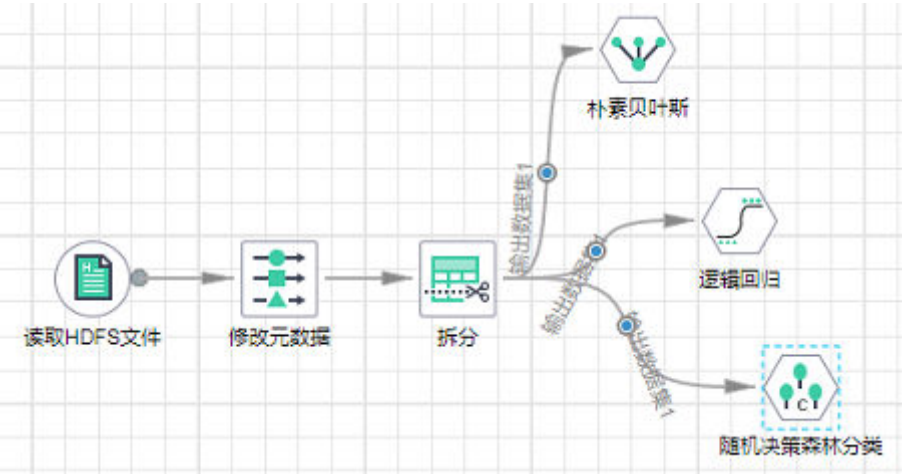
表 1-7 “拆分”节点参数配置样例

参数名称	样例值
分裂方式	比例
子数据集1的比例	66.0
随机种子	1

步骤5 模型训练。使用训练集分别训练朴素贝叶斯、逻辑回归和随机森林3种模型。操作如下：

1.
- 分别将“建模 > 分类”展开目录中的“朴素贝叶斯”节点、“逻辑回归”节点和“随机决策森林分类”节点拖拽至画布中。按图1-3将“拆分”节点分别和“朴素贝叶斯”节点、“逻辑回归”节点、“随机决策森林分类”节点连接。其中，设置“拆分”节点的“输出数据集1”为3个模型节点的输入。

图 1-3 连接各模型节点



2. 单击各节点，在右侧参数配置区域，并参照表1-8、表1-9和表1-10配置参数。

表 1-8 “朴素贝叶斯”节点参数配置

参数名称	样例值
平滑参数	1.0

表 1-9 逻辑回归”节点参数配置

参数名称	样例值
正则化函数	L1 and L2
正则化参数	0.001
弹性网络参数	0.5
最大迭代次数	100

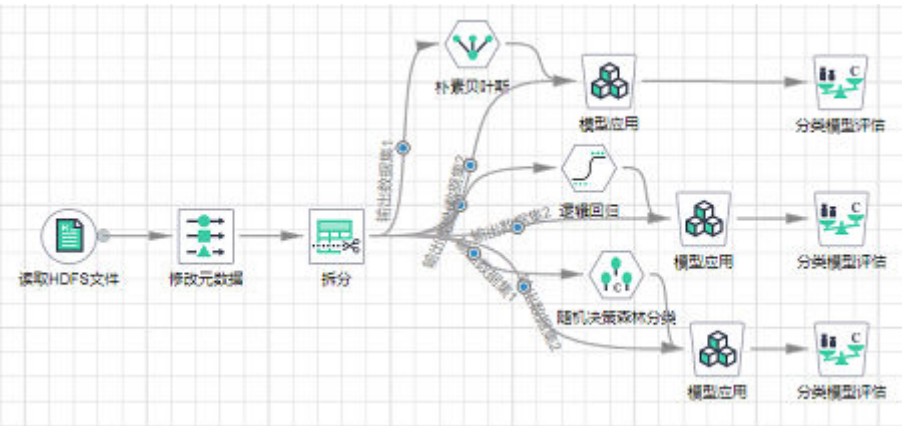
表 1-10 “随机决策森林分类”节点参数配置样例

参数名称	样例值
树的数目	100
最大树深度	4
最大分箱数	100
不纯度	Gini
特征子集选取策略	Auto
随机种子	1

步骤6 对3种模型添加性能评估指标。分别使用混淆矩阵、AUC和F1 score值对3个模型进行评估。操作步骤如下：

1. 将“评估”展开目录中的“模型应用”节点和“分类模型评估”节点拖拽至画布中，分别拖3个。参考图1-4链接各节点，设置“拆分”节点的“输出数据集2”为3个“模型应用”节点的输入，并将3个算法节点和“模型应用”相连，以及将“模型应用”和“分类模型评估”节点节点相连。


图 1-4 连接“模型应用”节点和“分类模型评估”节点



2. 单击各“模型应用”节点，在右侧参数配置区域都参照表1-11配置参数。

表 1-11 “模型应用”节点参数配置样例

参数名称	样例值
预测类型	分类
分类阈值	0.5，即预测值如果大于0.5样本为正样本，否则为负样本。

步骤7 单击，保存模型训练 workflow。

----结束

1.5 训练、评估和筛选模型

搭建好模型训练和评估的 workflow 后，接下来将运行 workflow，并得到各模型以及模型的评估结果。通过比较3个模型的评估结果，选取最佳性能的模型供后续业务分析使用。

操作步骤

步骤1 单击，运行 workflow。

步骤2 workflow 运行完毕后，分别右键单击3个“分类模型评估”节点，单击“查看评估结果”，3个模型的评估结果如图1-5、图1-6和图1-7（不同的 workflow 由于某些参数的随机性，会导致每次结果有一定差异）。

图 1-5 “朴素贝叶斯” 模型的评估结果

查看评估结果

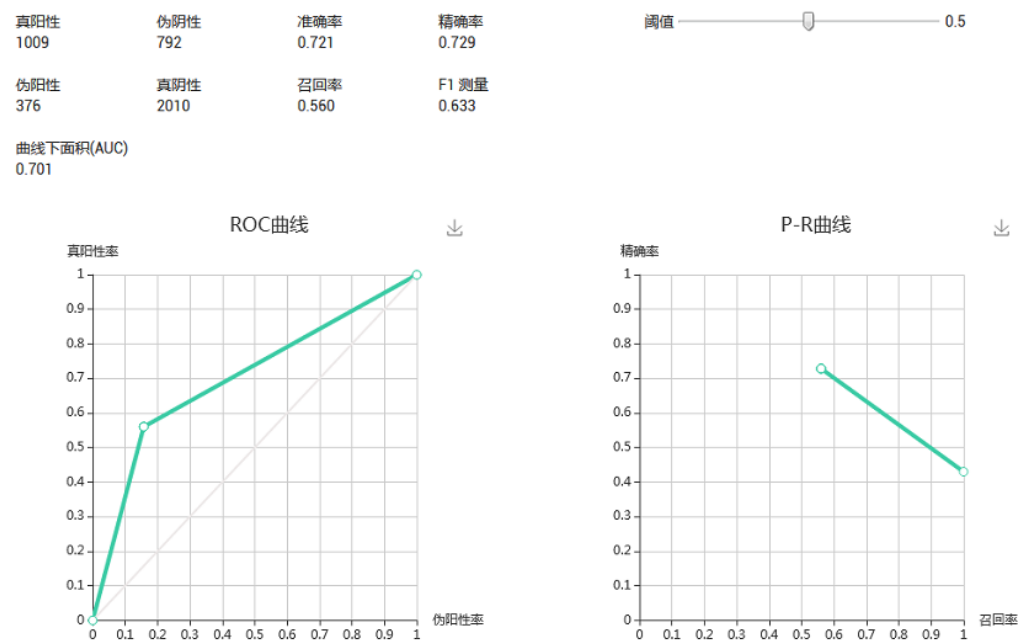


图 1-6 “逻辑回归” 模型的评估结果

查看评估结果

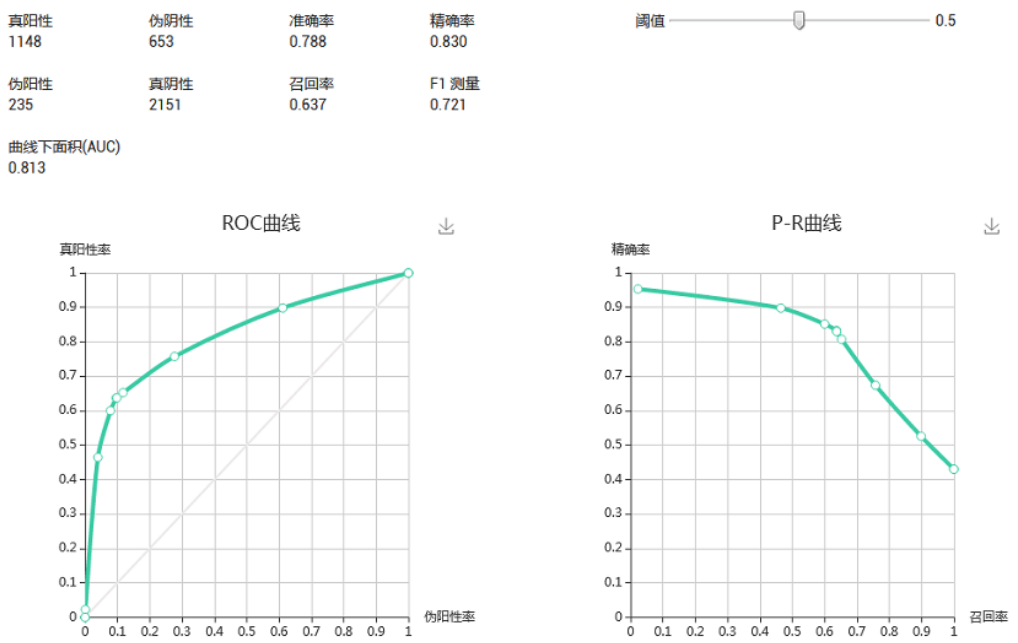
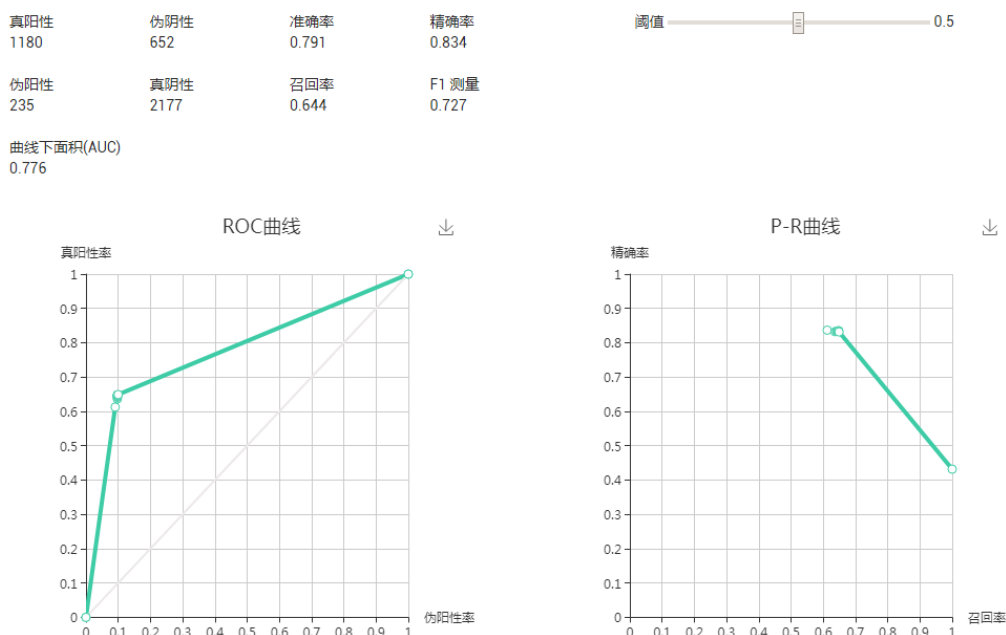


图 1-7 “随机森林”模型的评估结果

查看评估结果



步骤3 计算与分析各模型评估指标，如下：

- AUC
 - 朴素贝叶斯：0.701
 - 逻辑回归：0.813
 - 随机森林：0.776

AUC表示为ROC曲线下方的面积，简单来说，AUC值越大，说明模型分类正确率越高，详情参考[维基百科](#)。从3种模型的AUC值可以看出，逻辑回归模型分类准确率最好，朴素贝叶斯模型比较差。

- 精确率
 - 朴素贝叶斯：0.729
 - 逻辑回归：0.830
 - 随机森林：0.834

准确率表示样本中被识别成正样本准确率，即正样本被识别成正样本的个数与所有样本被识别成正样本个数比例，衡量模型的查准率，详情参考[维基百科](#)。从3种模型的精确率可以看出，随机森林模型识别正样本的准确率最高，朴素贝叶斯模型比较差。

- 召回率
 - 朴素贝叶斯：0.560
 - 逻辑回归：0.637
 - 随机森林：0.644

召回率表示样本中正样本被识别成正样本的比例，即被识别成正样本的个数与实际正样本的个数比例，衡量模型的查全率，详情参考[维基百科](#)。从3种模型的召回率可以看出，随机森林模型从正样本识别出正样本的比例最大，查全率最高。

- F1 score

- 朴素贝叶斯: 0.633
- 逻辑回归: 0.721
- 随机森林: 0.727

F1 score是统计学中用来衡量二分类模型精确度的一种指标。可以看作是模型准确率和召回率的一种加权平均, 详情参考[维基百科](#)。从3种模型的F1 score曲线图可以看出, 随机森林的值最大, 说明该模型的二分类精确度最高, 也最稳定。

综合比较模型的各项评估指标。最后, 选择“随机森林”作为本次业务分析的最终模型。

说明

对于对算法原理很了解的用户, 还可以进一步通过调整模型的参数, 使模型的性能更好。

----结束

1.6 保存模型

经过模型的训练和各模型的评估和筛选, 确认了使用“随机森林”模型作为本次业务分析的模型。接下来, 将模型保存在HDFS中, 便于后续使用时, 可以直接导入。

操作步骤

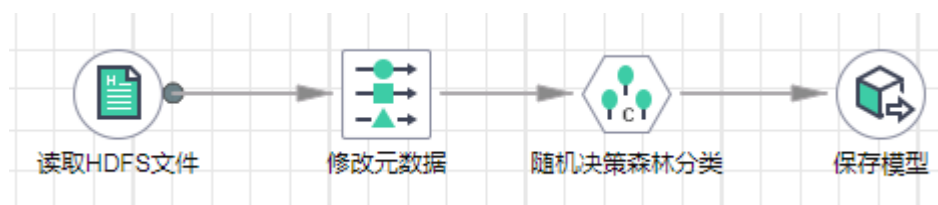
步骤1 将“输出”下拉目录中的“保存模型”节点拖入画布。单击“保存模型”节点, 在右侧参数配置区域按照[表1-12](#)配置参数。


表 1-12 “保存模型”节点参数配置样例

参数名称	样例值
模型路径	“/user/bank/output/”
模型文件名	RandomForest_bank
允许覆盖	勾选

步骤2 删除多余节点, 仅保留“读取HDFS文件”节点、“修改元数据”节点、“随机决策森林分类”节点和“保存模型”节点, 并按[图1-8](#)所示把各节点相链接。

图 1-8 输出模型



步骤3 单击 , 运行工作流。运行结束后, 模型将被保存到“/user/bank/output/”路径下。

----结束

1.7 搭建模型预测 workflow

得到“随机森林”模型后，我们将使用该模型对[业务场景介绍](#)中提到的3个新客户数据“NEW_DATA”进行分析，预测这两位客户是否具有存款意愿。此时，需要新建一个 workflow，并在 workflow 中导入新用户数据和随机森林模型，然后得到预测结果。

操作步骤

- 步骤1** 新建并打开一个 workflow，命名为“bank_predicting”，进入 workflow 编排界面。
- 步骤2** 导入新客户数据集。使用“读取HDFS文件”节点从HDFS中读取业务数据，操作步骤如下：

1. 将“输入”展开目录中的“读取HDFS文件”节点拖拽至画布中，单击该节点，在右侧参数配置区域按照[表1-13](#)配置参数。

表 1-13 “读取 HDFS 文件”节点参数配置样例

参数名称	样例值
数据格式	CSV
数据文件	“/user/omm/mls/NEW_DATA.csv”
导入元数据	不勾选
是否包含表头	勾选
字段分隔符	,
保存元数据文件	不勾选
处理异常值	null替代值
保存异常记录	不勾选

- 步骤3** 读取“随机森林”模型。使用“读取模型”节点导入模型，用于预测分析，操作如下：

1. 将“输入”展开目录中的“读取模型”节点拖拽至画布中，单击该节点，在右侧参数配置区域按照[表1-14](#)配置参数。

表 1-14 “读取模型”节点参数配置

参数名称	样例值
文件路径	“/user/bank/output/RandomForest_bank”

- 步骤4** 模型应用。使用“模型应用”节点对导入的数据集进行预测，操作如下：

1. 将“评估”展开目录中的“模型应用”节点拖拽至画布中，单击该节点，在右侧参数配置区域按照[表1-15](#)配置参数。

表 1-15 “模型应用” 节点参数配置

参数名称	样例值
预测类型	分类
分类阈值	0.5，即预测值如果大于0.5样本为正样本，否则为负样本。

步骤5 保存预测结果。使用“保存HDFS文件”节点将预测后的结果保存到HDFS中，操作如下：

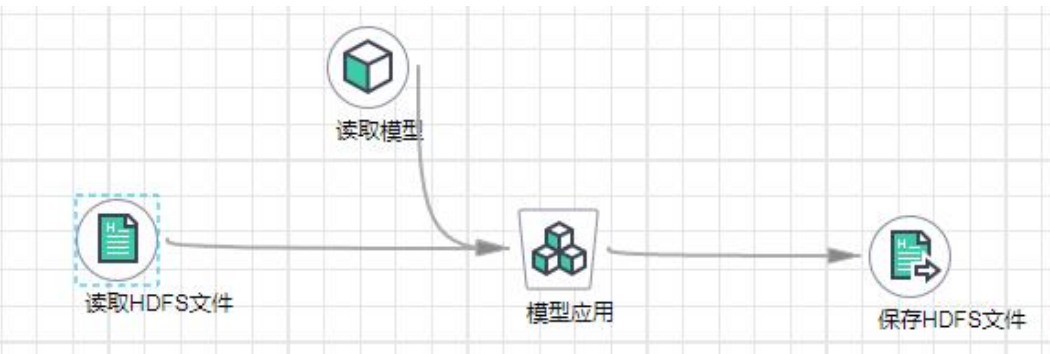
1. 将“输出”展开目录中的“保存HDFS文件”节点拖拽至画布中，单击该节点，在右侧参数配置区域参照表1-16配置参数。


表 1-16 “保存 HDFS 文件” 节点数说明

参数	样例值
文件目录	“/user/bank/output/”
文件名	predictData
文件格式	CSV
字段分隔符	,
允许覆盖	勾选

步骤6 并根据图1-9搭建模型预测的工作流。

图 1-9 读取模型到工作流



步骤7 单击，保存模型预测工作流。

----结束

1.8 预测数据

完成模型预测的工作流后，运行工作流即可得到预测结果。

操作步骤


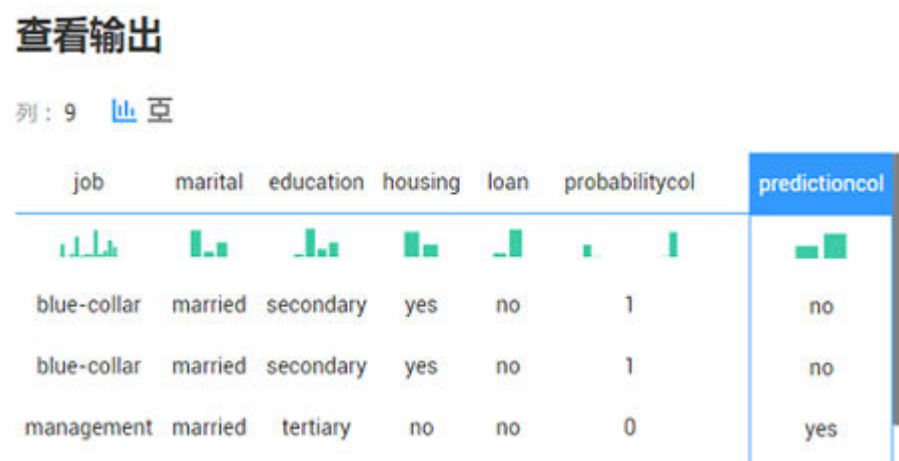
- 步骤1 单击，运行工作流。运行结束后，右键“模型应用”，单击“查看输出“输出数据集””。
- 步骤2 查看左侧列表中的“predicationcol”列的数据，即为预测结果，如图1-10

图 1-10 预测结果



- 步骤3 预测结果分析。从预测结果来看：
1.

客户1的信息为：年龄41岁，职业为蓝领，已婚，中等教育，有房贷，无个人贷款。预测结果为：该新客户无存款的意愿。
2.

客户2的信息为：年龄53岁，职业为蓝领，已婚，初级教育，有房贷，无个人贷款。预测结果为：该新客户无存款的意愿。
3.

客户3的信息为：年龄38岁，职业为管理，已婚，高等教育，无房贷，无个人贷款。预测结果为：该新客户有存款的意愿。

----结束

截止该步骤，即完成了如何使用机器学习，利用老客户数据，预测新客户的存款意愿。如后续再有新的客户数据得到，可以将数据从3.5章节的步骤2中导入，进行预测。

2商品批发商客户分群

使用 workflows 进行模型训练，获取客户分群的结果。

2.1 业务场景介绍

场景描述

在数据挖掘应用中，客户分群是一项重要的商业应用领域。通过对客户数据进行挖掘分析实现对客户做科学的分群。依据不同分群的特点制定相应的策略，从而为用户提供适配的产品、制定针对性的营销活动和管理用户，最终提升产品的客户满意度，实现商业价值。

本次业务场景为：某商品批发商手头有大量的客户数据，其中包括客户的个人信息以及每年到批发商采购各类商品的花销。商品批发商希望能通过这些客户数据，对客户进行分群，针对不同的客户群定制不同的营销策略。

数据说明

本次业务场景，我们使用来自UCI的“Wholesale customers Data Set”数据集模拟商品批发商手上的客户数据，下载地址为<http://archive.ics.uci.edu/ml/datasets/Wholesale+customers>。该数据集一共有440个样本，每个样本有8个特征数据，分别是：客户进货渠道、所在区域和不同商品（包括：生鲜类、奶制品、杂货、冷冻品、洗涤类和熟食类）的年度采购花销。我们把获取到的数据保存为“wholesale_customers_data_withTitle.csv”，同时往数据集中添加一列，列名为ID，从1开始按顺序取值，该ID用于客户身份识别，以区分不同客户的数据，部分数据如表2-1。

表 2-1 数据集部分样本数据

ID	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_papers	Delicatessen
1	2	3	12669	9656	7561	214	2674	1338
2	2	3	7057	9810	9568	1762	3293	1776
3	2	3	6353	8808	7684	2405	3516	7844

ID	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_papers	Delicatessen
4	1	2	13265	1196	421	6404	507	1788
5	2	4	22615	5140	7198	3915	1777	5185
6	2	3	9413	8259	5126	666	1795	1451
7	2	3	12126	3199	6975	480	3140	545

根据该数据集的相关描述，本次业务场景的问题可以理解为：

某商品批发商手头上有一份客户数据“wholesale_customers_data_withTitle.csv”，批发商希望将自己的客户分为3个客户群，分别是：

1. 大客户：大部分产品类别的年进货量都很大。
2. 中客户：大部分产品类别的年进货量都居中。
3. 小客户：大部分产品类别的年进货量都很小。

解决方案

客户分群问题属于数据挖掘中的聚类分析，我们从数据集数据中选择合适的特征，通过聚类算法完成客户的分群。在该业务场景中，用户的业务目标是基于客户在各产品类的消费情况进行分群，因此，渠道参数Channel和区域参数Region将不作为特征参与聚类分析。

聚类问题，就是给定一个元素集合D，其中每个元素具有n个可观察属性，使用某种算法将D划分成k个子集，要求每个子集内部的元素之间相异度尽可能低，而不同子集的元素相异度尽可能高。其中每个子集叫做一个簇。聚类算法非常多（几十种），其中，K-Means算法是最常用的一种。

K-Means是聚类算法中最常用的一种，算法最大的特点是简单，好理解，运算速度快，但是只能应用于连续型的数据，该算法要求在聚类前指定分成几类。K-Means聚类的目的：把n个样本点划分到k个分簇中，使得每个点离它所属分簇的分簇中心最近。算法的详细描述请参考[维基百科](#)。

在本次业务场景中，我们将使用K-Means算法对该数据集进行聚类分析。

2.2 操作流程概述

流程介绍

使用机器学习服务进行数据聚类的操作过程主要包括4个步骤：

1. [业务数据上传](#)。将业务数据上传通过OBS和MRS，上传到HDFS中，供用户分享使用。
2. [搭建模型训练工作流](#)。根据批发商手头的客户数据“wholesale_customers_data_withTitle.csv”，训练K-means聚类模型，此步骤主要是完成模型训练的工作流搭建操作。
3. [训练、评估和筛选模型](#)。运行工作流后得到K-means模型，并对模型的聚类中心进行评估分析，确认该模型的聚类性能。

4. **获取客户分群结果**。根据已训练的模型，再对客户数据进行聚类分析，得到最终的聚类结果。

2.3 业务数据上传


使用机器学习服务前，需要将本地数据文件上传至OBS，再通过MRS将数据从OBS中导入至HDFS，供MLS从HDFS中读取数据。

上传数据至 OBS

- 步骤1** 登录华为云管理控制台。
- 步骤2** 单击“服务列表”，选择“存储 > 对象存储服务”。
- 步骤3** 单击右上角的“创建桶”，进入“创建桶”页面。

表 2-2 “创建桶”样例

参数名	样例值
区域	选择新建桶所在的区域。
桶名称	obs-mls

- 步骤4** 单击“创建桶”，创建桶成功，返回“对象存储服务”页面。
- 步骤5** 在桶列表中，选择并单击“obs-mls”，进入“桶：obs-mls”界面。
- 步骤6** 在左侧导航栏中选择“对象”，单击“上传”，弹出“上传对象”页面。
- 步骤7** 单击，在弹出框中选择待上传的数据文件“wholesale_customers_data_withTitle.csv”。
- 步骤8** 单击“确定”，等待页面提示上传成功。

----结束

通过 MRS 将数据导入 HDFS

- 步骤1** 单击“服务列表”，选择“EI企业智能>MapReduce服务”。
- 步骤2** 在左侧导航栏中，选择“集群列表 > 现有集群”。
- 步骤3** 选择当前MLS实例所关联的MRS集群的名称，例如：“mrs-mls”。选择“文件管理”分页，单击“导入数据”。
- 步骤4** 在弹出框中，选择表2-3中的路径。

表 2-3 导入路径

路径	样例值
OBS路径	OBS中待导入的数据文件，例如“s3n://obs-mls/wholesale_customers_data_withTitle.csv”。
HDFS路径	数据文件导入到HDFS中的路径，例如“/user/omm/mls”。

步骤5 单击“确定”，等待页面提示导入成功。

----结束

2.4 搭建模型训练 workflow

操作场景

模型训练 workflow 主要包括4部分内容：

- 1. 读取业务数据集。
- 2. 数据集预处理，包括筛选特征和数据归一化处理。
- 3. 使用“K-均值”节点训练聚类模型。
- 4. 保存聚类模型。

操作步骤

- 步骤1** 新建一个 workflow，命名为“CustomerModel”。单击 workflow 名称，进入 workflow 编排界面。
- 步骤2** 读取数据。使用“读取HDFS文件”节点从HDFS中读取业务数据，操作步骤如下：
- 将“输入”展开目录中的“读取HDFS文件”节点拖拽至画布中。单击该节点，在右侧参数配置区域按照表2-4配置参数。

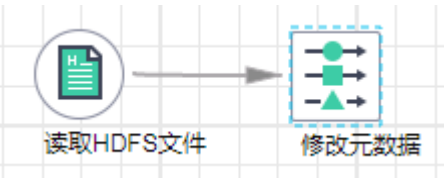
表 2-4 “读取 HDFS 文件”节点参数配置样例

参数名称	样例值
数据格式	CSV
数据文件	“/user/omm/customer/wholesale_customer_data_withTitle.csv”
导入元数据	不勾选
是否包含表头	勾选
字段分隔符	“，”
保存元数据文件	不勾选
处理异常值	null替代值
保存异常记录	不勾选

步骤3 筛选模型训练的特征。在实际使用中，数据集的第1列ID、第2列Chanel和第3列Region不参与聚类分析，因此需要剔除。使用“修改元数据”节点将这3列设置为不参与训练模型特征列，操作步骤如下：

1.
- “数据转换 > 字段操作”展开目录中的“修改元数据”节点拖拽至画布中。如图2-1将“读取HDFS文件”节点和“修改元数据”节点相连。

图 2-1 连接“修改元数据”节点



2.
- 单击该节点，在右侧参数配置区域按照表2-5配置“字段”和“角色”，其他参数保持默认。

表 2-5 “修改元数据”节点参数配置样例

字段	角色	说明
id	None	客户ID值，不作为特征值。
channel	None	渠道参数，不作为特征值。
region	None	区域参数，不作为特征值。
fresh	Input	特征值
milk	Input	特征值
grocery	Input	特征值
frozen	Input	特征值
detergents_paper	Input	特征值
delicatessen	Input	特征值

步骤4 数据归一化。通过观察数据集统计数据值，如表2-6，各特征列的数值范围差异较大，如果直接用原始值进行计算，会出现取值范围大的属性对距离的影响高于取值范围小的属性，这样不利于反映取值范围小的特征的真实相异度，因此，我们将这六个特征值按比例映射到相同的[0,1]区间，从而平衡各个特征对距离的影响。

表 2-6 数据集特征信息

编号	特征名称	最小值	最大值	均值	标准方差
1	Fresh	3	112151	12000.30	12647.329

编号	特征名称	最小值	最大值	均值	标准方差
2	Milk	55	73498	5796.27	7380.377
3	Grocery	3	92780	7951.28	9503.163
4	Frozen	25	60869	3071.93	4854.673
5	Detergents_papers	3	40827	2881.49	4767.854
6	Delicatessen	3	47943	1524.87	2820.106

使用“标准化”节点实现数据的归一化处理，操作步骤如下：

1. 将“数据转换 > 字段操作”展开目录中的“标准化”节点拖拽到画布中。如图2-2将“标准化”节点和“修改元数据”节点相连。

图 2-2 连接“标准化”节点



2. 单击该节点，在右侧参数配置区域按照表2-7配置参数。

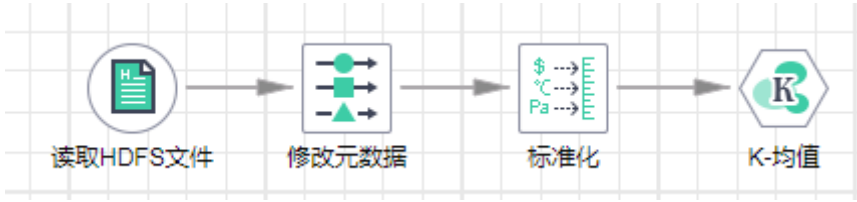
表 2-7 “标准化”节点参数配置样例

参数名称	样例值
特征	选择所有特征参数
方法	极差变换
最小值	0
最大值	1

步骤5 模型训练。使用“K-均值”节点训练聚类模型，操作步骤如下：

1. 将“建模 > 聚类”展开目录中的“K-均值”节点拖拽到画布中。如图2-3将“K-均值”节点和“标准化”节点相连。


图 2-3 连接“K-均值”节点



2. 单击该节点，在右侧参数配置区域按照表2-8配置参数。

表 2-8 “K-均值” 节点参数配置样例

参数名称	样例值
聚类数	3
迭代次数	20
初始模式	K-MeansII
初始模式步数	5

步骤6 单击，保存模型训练工作流。

----结束

2.5 训练与评估模型

搭建好模型训练的工作流之后，接下来将进行运行工作流，得到聚类模型，通过对聚类模型的可视化结果和聚类中心的分析，评估模型的聚类性能。

操作步骤


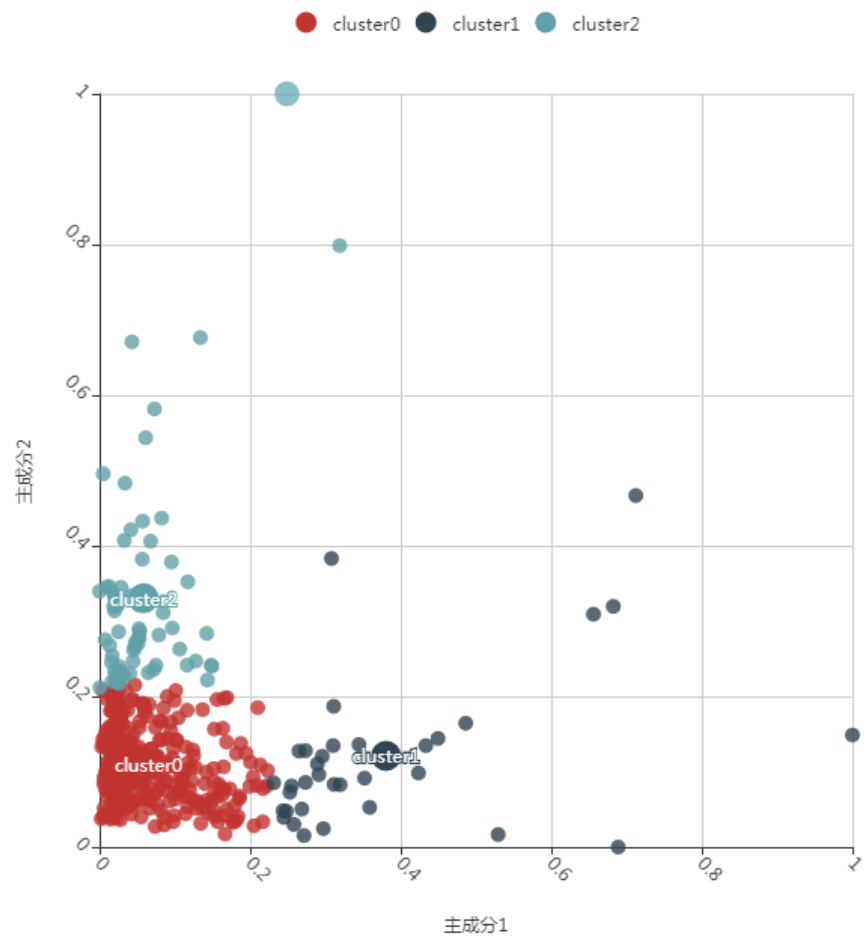
- 步骤1** 单击，运行工作流。
- 步骤2** 当工作流运行结束后，右键“K-均值”节点，单击“查看模型”，弹出查看模型界面，如图2-4所示。

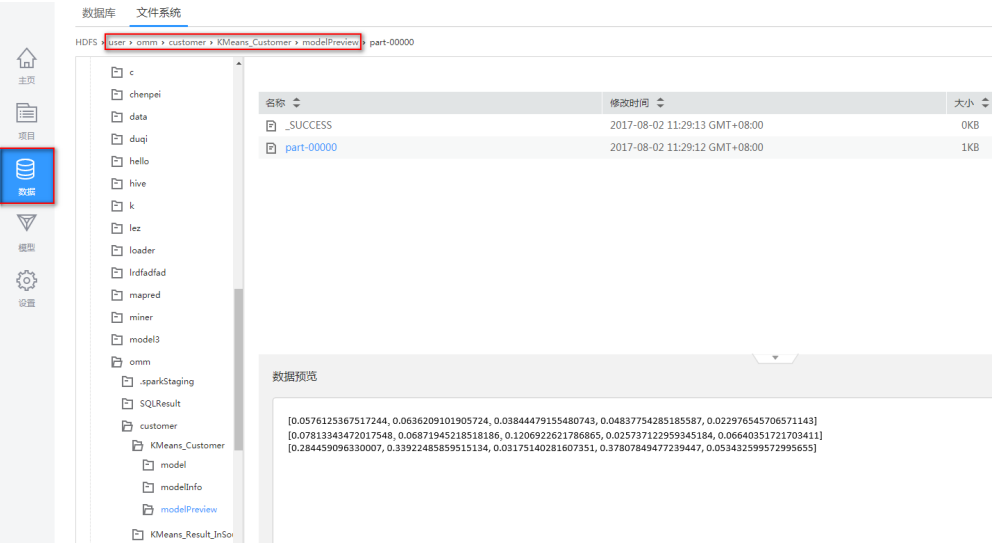
图 2-4 聚类模型的可视化效果



从图中可以看出，数据集通过降维处理后得到2维主成分。所有样本都被较好的划分到了3类中的某类，样本间基本无交叉现象，3个类别的聚类中心相聚一定距离。从图2-4中看出，该聚类效果较为理想。

步骤3 在实例工作页面，单击项目页面的“数据”页，进入到存放结果文件的路径，查看“保存模型”节点的输出文件，如图2-5所示。

图 2-5 “K-均值”模型分簇中心结果



聚类中心点数值如表2-9所示，通过聚类中心点结果分析客户分群情况。

表 2-9 归一化后的聚类中心点数值

簇	Fresh	Milk	Grocery	Frozen	Detergents_papers	Delicatesen
0	0.07135554117480078	0.0576125367517244	0.0636209101905724	0.03844479155480743	0.04837754285185587	0.022976545706571143
1	0.30708185063144583	0.07813343472017548	0.06871945218518186	0.1206922621786865	0.025737122959345184	0.06640351721703411
2	0.08169063359552188	0.284459096330007	0.33922485859515134	0.03175140281607351	0.37807849477239447	0.053432599572995655

在表2-9中：

1. 分簇2中心点各特征值维度上的值分别为分簇1中心点的0.27、3.6、4.9、0.26、14.7和0.8倍。
2. 分簇1中心点各特征值维度上的值又分别为分簇0中心点的4.3、1.36、1.08、3.14、0.53和2.89倍。

可知，分簇2中心点的值总体上大于分簇1的中心点，而分簇1中心点的值又基本上全面大于分簇0的中心点。所以，我们可确认分簇2为大客户分群、分簇1为中客户分群、分簇0为小客户分群。

----结束

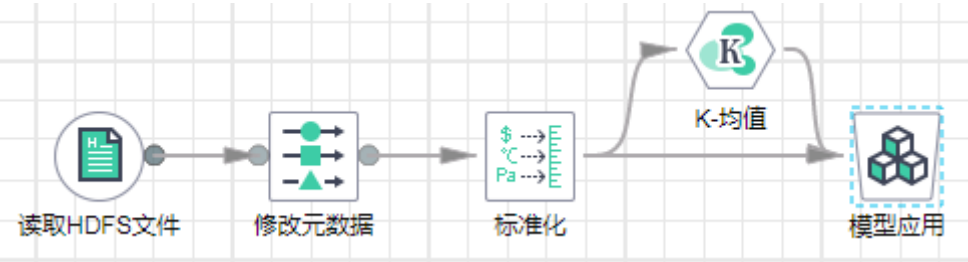
2.6 获取客户分群结果

通过分析聚类模型的聚类性能和聚类分簇所代表的客户群体，我们需要分析批发商手中所有客户的分群情况。

操作步骤

- 步骤1** 使用聚类模型。使用“模型应用”节点对业务数据进行分类操作。操作步骤如下：
- “评估”展开目录中的“模型应用”节点拖拽到画布中。如图2-6将“模型应用”节点和“K-均值”节点和“标准化”节点相连接。

图 2-6 连接“模型应用”节点



- 单击该节点，在右侧参数配置区域按照表2-10配置参数。

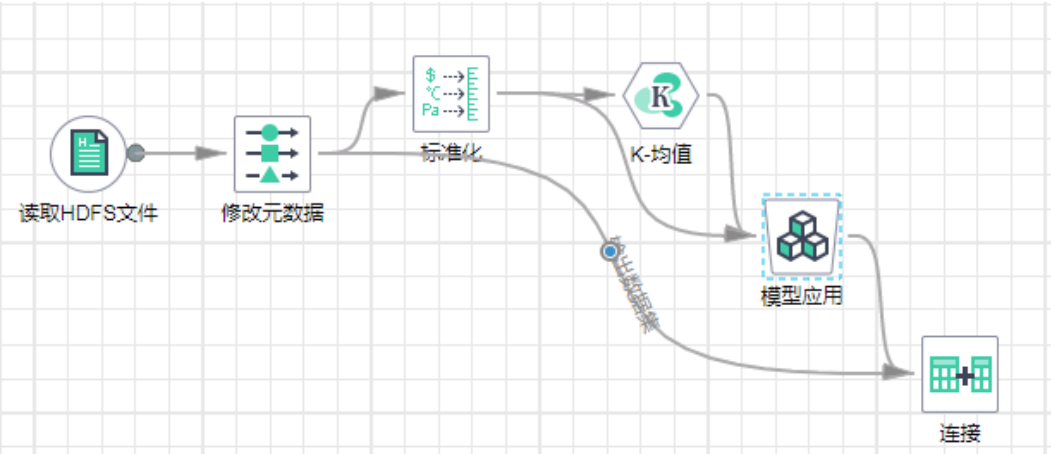
表 2-10 “模型应用”节点参数配置样例

参数名称	样例值
预测类型	聚类

- 步骤2** 由于数据集进行了标准化处理，因此从“模型应用”出来的结果，每个样本的数据都变了。因此，我们可以通过“连接”把“模型应用”输出的聚类结果和标准化前的业务数据进行连接，得到原始数据的聚类结果。操作步骤如下：

- 将“数据转换 > 记录操作”展开目录中的“连接”节点拖拽到画布中。将“修改元数据”节点和“连接”节点连接。“修改元数据”节点的输出数据集设置为“连接”节点的左输入数据集。将“模型应用”节点和“连接”节点连接。“模型应用”节点的输出会自动作为“连接”节点的右输入数据集，如图2-7所示。

图 2-7 连接“连接”节点



2.
- 单击该节点，在右侧参数配置区域按照表2-11配置参数。

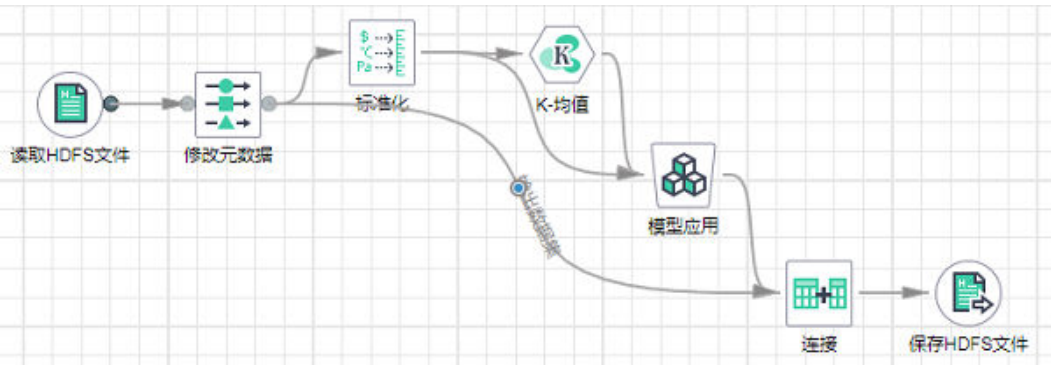
表 2-11 “连接”节点参数配置样例

参数名称	样例值
连接类型	Inner
关联字段	单击  ，增加关联字段，并把“左表字段”和“右表字段”设为“id”。
合并关联字段	勾选
合并重名的字段	勾选

步骤3 保存聚类结果。使用“保存HDFS文件”节点将聚类结果保存起来，操作步骤如下：

1.
- 将“输出”展开目录中的“保存HDFS文件”节点拖拽到画布中。将“连接”节点和“保存HDFS文件”节点连接，如图2-8所示。

图 2-8 连接“保存 HDFS 文件”节点



2.
- 单击“保存HDFS文件”节点，在右侧参数配置区域按照表2-12配置参数。

表 2-12 “保存 HDFS 文件” 节点参数配置样例

参数名称	样例值
文件目录	保存标准化结果文件的路径
文件名	KMeans_Result_InSourceData
文件格式	CSV
字段分隔符	“ ” ,
允许覆盖	勾选


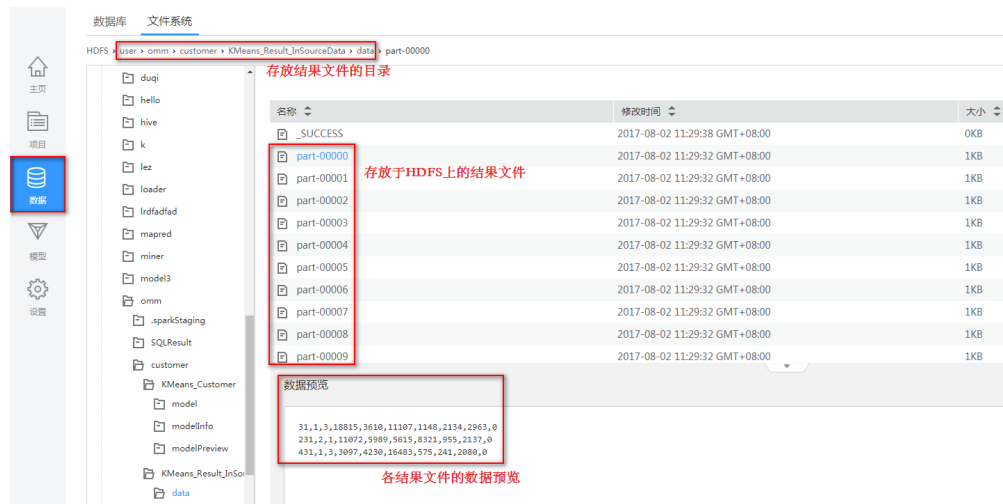
- 步骤4 单击，运行工作流。
- 步骤5 当工作流运行结束后，返回实例工作界面，单击左侧导航栏的“数据”，进入“数据”页面。选择“文件系统”页签，进入存放结果文件的路径，查看聚类分簇结果的输出文件，如图2-9所示。

图 2-9 聚类分簇结果



从客户分群结果中为每个客户分群选取10个分群结果文件，部分客户分群结果如表 2-13。

表 2-13 客户样本分群结果

ID	Fresh	Milk	Grocery	Frozen	Detergents_papers	Delicate ssen	客户分群
44	630	11095	23998	787	9529	72	大客户
46	5181	22044	21531	1740	7353	4985	大客户
48	44466	54259	55571	7782	24671	6465	大客户
50	4967	21412	28921	1798	13583	1163	大客户

ID	Fresh	Milk	Grocery	Frozen	Detergents_papers	Delicatessen	客户分群
57	4098	29892	26866	2616	17740	1340	大客户
252	6134	23133	33586	6746	18594	5121	大客户
34	29729	4786	7326	6130	361	1083	中客户
37	29955	4362	5428	1729	862	4626	中客户
41	24025	4332	4757	9510	1145	5864	中客户
233	25962	1780	3838	638	284	834	中客户
241	22096	3575	7041	11422	343	2564	中客户
437	39228	1431	764	4510	93	2346	中客户
31	18815	3610	11107	1148	2134	2963	小客户
33	21632	1318	2886	266	918	405	小客户
231	11072	5989	5615	8321	955	2137	小客户
237	8635	820	3047	2312	415	225	小客户
431	3097	4320	16483	575	241	2080	小客户
433	21117	1162	4754	269	1328	395	小客户

在表2-13中可以看出。

1. 大客户在Milk、Grocery和Detergents_papers分类上的年进货花销全面大于中客户，在Frozen和Delicatessen分类上则没有太大的差别，仅在Fresh分类上大部分小于中客户的年进货花销。在总年进货花销上，大客户明显领先于中客户。
2. 中客户在Fresh分类上的年进货花销大于小客户，在其余分类上没有太大的差别。在总年进货花销上，中客户明显领先于小客户。

由上述分析可知，通过“K-均值”聚类算法，能很好地实现了对批发商的客户进行大客户、中客户和小客户的分群。

----结束

3 电影推荐

使用 workflow 进行模型训练并保存模型文件，再使用 workflow 实现客户电影推荐。

3.1 业务场景介绍

场景描述

近几年，随着电影技术的发展，电影已经成为人们日常消遣娱乐放松的主要方式之一，人们在观看完电影后，通常会对一部电影进行评价，因此会产生大量的和电影评价相关的数据集。那么如何对已有的数据进行分析，向用户准确推荐他们感兴趣的电影，正是我们需要解决的问题。

本次业务场景为：某电影票团购公司有大量观众对电影评分的数据。该公司希望利用这些历史数据，分析某些用户对未做评分的电影感兴趣程度（评分），根据分析结果（评分等级），向感兴趣用户准确推荐新上线的电影。

数据说明

本次业务场景的模拟数据来自MovieLens，地址为<https://grouplens.org/datasets/movielens>。数据集包含了51093名用户对其中29200名电影的评价数据，共计646675条数据（该数据集还在不断的扩充当中），部分样本数据如表3-1。数据集包含3列特征，分别是user（观众ID）、item（电影ID）和rating（评分等级）。其中，评分等级分为0至10分，11个等级，相关的元数据结构如表3-2。

表 3-1 部分样本数据

user	item	rating
461	996	2
996	999	3
299	18	2
191	19	4
601	18	3

user	item	rating
99	19	4
692	18	4
216	19	2
531	524	2

表 3-2 元数据结构

名称	类型	衡量	值	角色
user	Integer	Continuous	[1,51093]	Input
item	Integer	Continuous	[8,7359702]	Input
rating	Real	Nominal	[0,10]	Target

根据业务场景描述，我们对数据集做如下处理：

- 从上述数据集中选取5个样本作为待分析的数据集。该数据集包含5个用户ID，某个电影ID（ID为417），和评分结果rating。数据集如表3-3，并数据为“new_data.csv”。
- 剩余的样本构成的数据集，作为团购公司手头上已有的历史客户数据，并数据为“movie_ratings.csv”。

表 3-3 待预测数据集

user	item	rating
37847	417	10
44672	417	8
38298	417	7
19113	417	7
29554	417	7

那么，该业务场景的问题可理解为：

某电影团购公司手头上有大量历史客户数据“movie_ratings.csv”，他们希望利用机器学习服务，预测某5位观众对该电影（ID为417）的评分。并对评分大于8的观众，推送该电影上线信息。

解决方案

通过问题描述可知这是一个电影的推荐问题，因此可以用“交替最小二乘”算法来分析处理：

- 交替最小二乘（Alternating Least Squares）

在机器学习中，ALS 指使用交替最小二乘求解的一个协同推荐算法。该方法常用于基于矩阵分解的推荐系统中。它通过观察到的所有用户给电影的打分，来推断每个用户的喜好并向用户推荐适合的电影。可以使用MLS创建包含“交替最小二乘”算子的工作流进行分析处理，最终输出模型，进行保存。

3.2 操作流程概述

使用机器学习服务进行推荐分析的操作过程主要包括5个步骤：

1. **业务数据上传**。将业务数据上传通过OBS和MRS，上传到HDFS中，供用户分享使用。
2. **搭建模型训练工作流**。根据电影票团购公司的客户数据“movie_ratings.csv”，训练推荐模型，此步骤主要是完成模型训练的工作流搭建操作。
3. **训练模型**。运行工作流后得到并保存交叉最小二乘模型。
4. **搭建模型预测工作流**。使用得到的模型根据某观众和电影数据“new_data.csv”，预测观众对电影评估值，此步骤主要是完成模型预测的工作流搭建操作。
5. **获取评分结果**。运行工作流，得到评分结果并对评分结果做简要说明。

3.3 业务数据上传

使用机器学习服务前，需要将本地数据文件上传至OBS，再通过MRS将数据从OBS中导入至HDFS，供MLS从HDFS中读取数据。

上传数据至 OBS

步骤1 登录公有云管理控制台。

步骤2 单击“服务列表”，选择“存储>对象存储服务”。


步骤3 单击“创建桶”。

填写以下参数配置样例，具体请参见“对象存储服务>用户指南>快速入门>管理控制台快速入门>创建桶”。

表 3-4 “创建桶”样例

参数名	样例值
区域	选择新建桶所在的区域。
桶名称	obs-mls

步骤4 单击“确定”，创建桶成功。

步骤5 选择桶“obs-mls”，单击，在弹出框中选择待上传的数据文件，确定后，单击“上传”，等待页面提示上传成功。

具体请参见“对象存储服务 > 用户指南 > 快速入门 > 管理控制台快速入门 > 创建桶”。

----结束

通过 MRS 将数据导入 HDFS

步骤1 单击“服务列表”，选择“EI企业智能 > MapReduce服务”，打开MRS控制台页面。

步骤2 选择集群“mrs-mls”，单击“文件管理 > 导入”。

步骤3 在弹出框中，选择表3-5中的路径。

表 3-5 导入路径

路径	说明
OBS路径	“s3n://obs-mls/movie_ratings.csv” 和 “s3n://obs-mls/new_data.csv”
HDFS路径	例如 “/user/zhongce”

步骤4 单击“导入”，等待页面提示导入成功。

----结束

3.4 搭建模型训练 workflow

操作步骤

步骤1 新建一个工作流，命名为“ALS_Movie”。单击工作流名称，并进入的工作流编排界面。

步骤2 读取数据。使用“读取HDFS文件”节点从HDFS中读取业务数据，操作步骤如下：

将“输入”展开目录中的“读取HDFS文件”节点拖拽至画布中。单击该节点，在右侧参数配置区域按如表3-6配置参数。

表 3-6 “读取 HDFS 文件” 节点参数配置样例

参数名称	样例值
数据格式	CSV
数据文件	“/user/zhongce/movie_ratings.csv”
导入元数据	不勾选
是否包含表头	勾选
字段分隔符	,

参数名称	样例值
保存元数据文件	不勾选
处理异常值	Replace with null
保存异常记录	不勾选

步骤3 模型训练。使用“交替最小二乘”节点训练推荐模型，操作步骤如下：

1. 将“建模 > 推荐”下拉目录中的“交替最小二乘”节点拖拽至画布中。单击该节点，在右侧参数配置区域按如表3-6表3-7配置参数。

表 3-7 “交替最小二乘”节点参数配置样例

参数名称	样例值
迭代次数	10
用户列名称	user
电影列名称	item
评分列名称	rating

2. 将“读取HDFS文件”节点连接到“交替最小二乘”节点，如图3-1。

图 3-1 连接“交替最小二乘”节点



步骤4 保存模型。使用“保存模型”节点将推荐模型保存起来，操作步骤如下：

1. 将“输出”展开目录中的“保存模型”节点拖入画布。单击“保存模型”节点，按照表3-8配置参数。


表 3-8 保存推荐模型”节点参数配置样例

参数名称	样例值
模型路径	“/usr/temp”
模型文件名	als_rating
允许覆盖	勾选

2. 将“保存模型”节点连接到“交替最小二乘”节点，如图3-2所示。


图 3-2 连接“保存模型”



步骤5 单击 ，保存模型训练工作流。
----结束

3.5 训练模型

操作步骤

步骤1 单击 ，运行工作流。
步骤2 待工作流运行结束后，可以在“/usr/temp”目录下查看模型文件。
----结束

3.6 搭建模型预测工作流

得到推荐模型后，我们将使用该模型对[业务场景介绍](#)中提到的5个客户数据“new_data”进行分析，预测客户对电影的评分。此时，需要新建一个工作流，并在工作流中导入新用户数据和推荐模型，然后得到预测结果。

操作步骤

步骤1 新建并打开一个工作流，命名为“movie_predicting”，进入工作流编排界面。
步骤2 导入新客户数据集。使用“读取HDFS文件”节点从HDFS中读取业务数据，操作步骤如下：

1. 将“输入”展开目录中的“读取HDFS文件”节点拖拽至画布中，单击该节点，在右侧参数配置区域按照[表3-9](#)配置参数。

表 3-9 “读取 HDFS 文件”节点参数配置样例

参数名称	样例值
数据格式	CSV
数据文件	“/user/zhongce/new_data.csv”
导入元数据	不勾选
是否包含表头	勾选
字段分隔符	,
保存元数据文件	不勾选

参数名称	样例值
处理异常值	null替代值
保存异常记录	不勾选

步骤3 读取推荐模型。使用“读取模型”节点导入模型，用于预测分析，操作如下：

1. 将“输入”展开目录中的“读取模型”节点拖拽至画布中，单击该节点，在右侧参数配置区域按照表3-10配置参数。

表 3-10 “读取模型”节点参数配置

参数名称	样例值
文件路径	“/user/temp/als_rating”

步骤4 模型应用。使用“模型应用”节点对导入的数据集进行预测，操作如下：

1. 将“评估”展开目录中的“模型应用”节点拖拽至画布中，单击该节点，在右侧参数配置区域按照表3-11配置参数。

表 3-11 “模型应用”节点参数配置

参数名称	样例值
预测类型	推荐
推荐商品个数	3

步骤5 保存预测结果。使用“保存HDFS文件”节点将预测后的结果保存到HDFS中，操作如下：

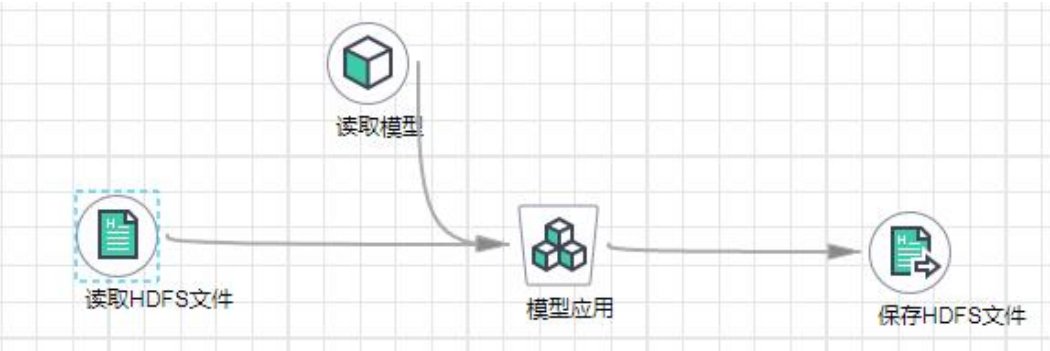
1. 将“输出”展开目录中的“保存HDFS文件”节点拖拽至画布中，单击该节点，在右侧参数配置区域参照表3-12配置参数。


表 3-12 “保存 HDFS 文件”节点数说明

参数	样例值
文件目录	“/user/bank/output/”
文件名	predictMovie
文件格式	CSV
字段分隔符	,
允许覆盖	勾选

步骤6 并根据图3-3搭建模型预测的工作流。

图 3-3 读取模型到工作流



步骤7 单击 ，保存模型预测工作流。
----结束

3.7 获取评分结果

完成模型预测的工作流后，运行工作流即可得到预测结果。

操作步骤





- 步骤1**
- 单击 ，运行工作流。运行结束后，右键“模型应用”，单击“查看输出“输出数据集””。
- 步骤2**
- 查看左侧列表中的“prediction”列的数据，即为预测结果，如图3-4

图 3-4 预测结果

user	item	rating	prediction
			
37847	417	10	8.762109
44672	417	8	8.270552
38298	417	7	7.6151476
19113	417	7	7.214668
29554	417	7	7.4375563

步骤3 预测结果分析。从图3-4可知，“prediction”列为模型预测的结果，“rating”列真实值，根据之前业务要求，团购公司将对评分结果大于8的观众推送该电影。那么，根据预测结果，公司将为ID为37847和ID为44672的观众推送该电影的相关信息。比较

“rating”列和“prediction”列，可以看出预测结果和真实评分的差异较小（除第一列差1.3左右），对于公司的业务需求（对评分大于8的用户推荐电影）并没有影响。由此可见，该模型能较准确的实现电影推荐业务。

----结束

A 附录

A.1 修订记录

发布日期	修改说明
2018-01-03	第一次正式发布。