**基于PySpark的客户流失分析研究**

**作者**：程合晴，王琛光，范冰冰，冯永

**单位：** 重庆大学，计算机学院

**案例版权：**该案例归重庆大学计算机学院所有

**涉及的知识点：**客户流失分析、二分类预测、逻辑回归、梯度提升树、超参调优、交叉验证、可视化、PySpark

**案例来源及案例真实性情况：**该案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。

**摘要** 客户流失率是指客户的流失数量与全部消费产品或服务客户的数量的比例， 它是客户流失的定量表述，是判断客户流失的主要指标，直接反映了企业经营与 管理的现状。客户流失分析可以更好地了解客户流失的根本原因，以减少客户流 失和提高产品销售额。因此，客户流失率分析是一个非常有研究价值的课题。本案例的核心内容为：（1）采用逻辑回归、梯度提升树方法进行用户流失的二分类预测，并使用AUC指标和PR指标进行评估。（2）使用K-Fold交叉验证进行模型超参调优。（3）绘制特征与流失用户关系图、特征图重要度实现结果可视化。

**关键词：**客户流失，分类预测，超参调优，可视化，PySpark

## 1 引言

该教学案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。该案例的关键问题为客户流失预测，通过分析，企业可以找出是什么原因导致客户离开，谁有可能离开，以及自己能做些什么来留住客户。基于以上，需引导学生进行的主要内容有：（1）采用逻辑回归、梯度提升树进行用户流失的二分类预测，并使用AUC指标、PR指标进行模型评估。（2）使用K-Fold交叉验证进行模型超参调优。（3）绘制特征与流失用户关系图、特征图重要度实现结果可视化。

## 2 背景介绍

随着社会经济的发展，企业之间的竞争压力日益加剧，许多公司都面临客户流失问题。越来越多公司意识到：挽留现有客户比吸引新客户更重要，因为成本更低，挽留现有客户是最好的市场营销方式。对于客户流失，一般分为两类：一类是意外流失，即客户发生财政状况变化等而导致被迫放弃服务；另一类是主动流失，即客户由于其他公司提供更优质的服务而决定将其服务转移。因此对于主动流失的客户，在流失之前进行预测，并采取适当营销策略挽留客户，有利于提高企业利润，也能避免企业在核心竞争中被淘汰。

本案例使用PySpark进行实现，PySpark提供了相关机器学习库MLlib 。MLlib主要包括三个主要的抽象类：转换器（Transformer）、评估器（Estimator）和管道（Pipeline）。其中，转换器通过将一个新列附加到DataFrame来转换数据，常见的转换包括二值化、阈值化、IDF词频统计等；评估器可以被视为需要评估的统计模型，对观测的对象做预测或分类，通过选用不同的评估器可以实现分类、回归、预测等功能；管道则表示从转换到评估（具有一系列不同阶段）的端到端的过程，这个过程可以对输入的一些原始数据以DataFrame形式执行必要的数据加工与转换，最后评估模型。

**3 内容**

本案例的主要内容分为三个部分，分别为二分类预测、超参调优与模型评估、可视化。本案例的实现语言为PySpark，所使用的数据集相关说明如下表所示。

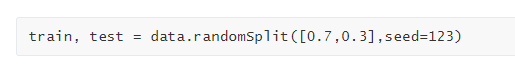
|  |  |
| --- | --- |
| 列名称 | 含义 |
| Names | 客户姓名 |
| Age | 客户年龄 |
| Total\_Purchase | 购买金额 |
| Account\_Manager | 有无客户经理跟进 |
| Years | 客户总年数 |
| Num\_Sites | 使用公司服务的网站数量 |
| Onboard\_date | 公司的多年用户 |
| Location | 客户地址 |
| Company | 客户公司 |
| Churn | 是否流失 |

3.1 用户流失分类预测与模型评估

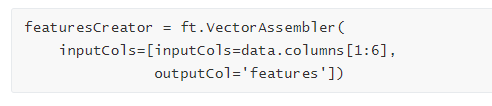
在本小节，我们划分数据集7：3作为训练集和测试集，分别训练逻辑回归和梯度提升树模型，实现客户流失的分类预测。同时，我们使用AUC指标和PR指标进行分类预测结果评估。使用PySpark实现该部分的具体步骤如下：

1. 数据预处理步骤：读取数据集，数据集按7：3划分训练集和测试集。

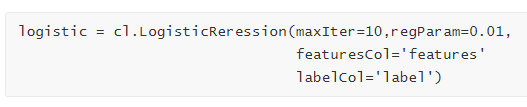




（2）特征工程步骤：为了进行回归预测，使用PySpark的VectorEmsseble作为转换器进行多列合并成向量，以便后续输入。

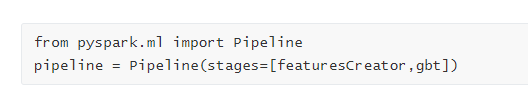


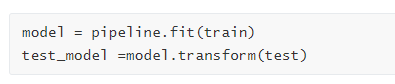
1. 模型训练步骤：构建逻辑回归和梯度提升树模型的评估器。



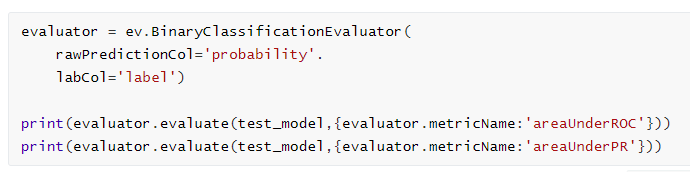


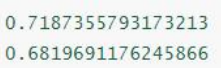
将之前构造的转换器和新构造的评器一起构建管道，下图以梯度提升树为例。

 使用管道进行模型拟合。



（4）模型评估步骤：使用拟合后的模型在测试集上进行测试，并使用AUC指标和PR指标进行评估。

得到结果如下，逻辑回归方法的AUC指标和PR指标：



梯度提升树方法的AUC指标和PR指标：

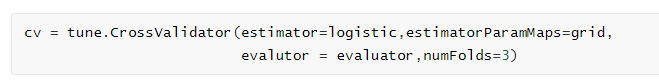


可以看出梯度提升树方法性能优于逻辑回归方法。

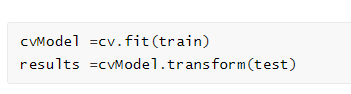
3.2 超参调优

第一个模型往往由于参数不是最佳的，所以通常不是最好的模型。通过使用超参调优能帮我们找到模型的最佳参数，如逻辑回归模型所需的最佳迭代次数或决策树的最大深度。在超参调优时，常用的验证方法有：K-Fold交叉验证和1-Fold交叉验证。在这里我们选取K-Fold交叉验证进行实验，该部分具体步骤如下：

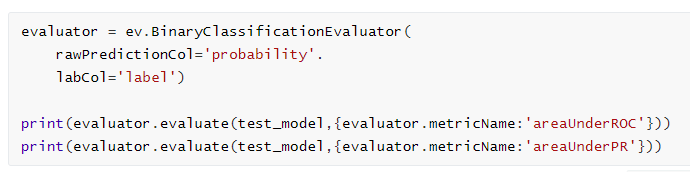
1. 使用梯度提升树方法进行训练，并使用K-Fold交叉验证评估各种参数模型，这里K取3。



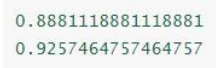
1. 返回估计的最佳模型



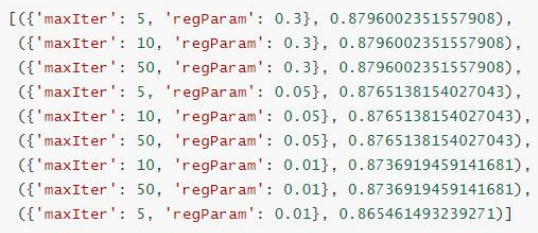
1. 再次进行模型评估，



得到ROC指标和PR指标结果如下：

 可以看出经过K-Fold交叉验证进行超参调优的模型，在ROC指标和PR指标上性能均有显著提升。

1. 以ROC指标为例，对得到模型最优参数进行输出查看，如下图所示。可以看出最优参数组合为｛maxIter=5,regParam=0.3}。



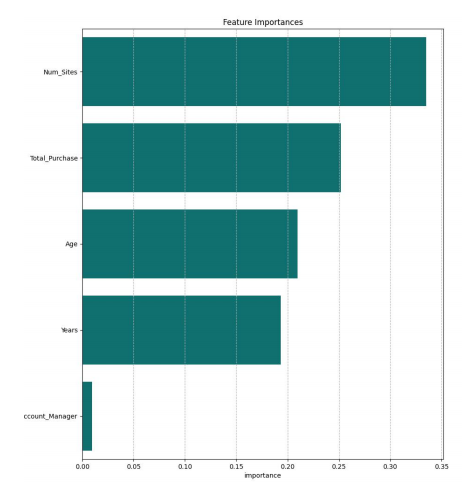
3.3 可视化展示

在这一部分，我们分别绘制了特征与流失用户关系图以及特征重要图。

1. 特征与流失用户关系图

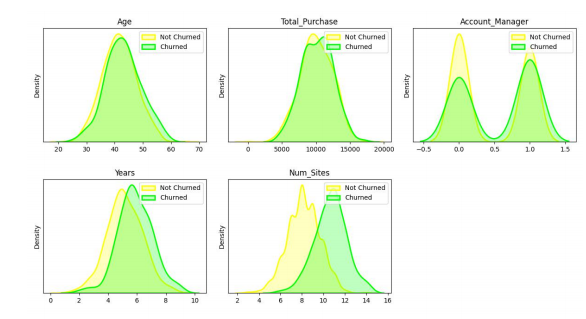
特征与流失用户关系图反映了流失客户和未流失客户在不同特征的分布，如下图所示。从图中可以看出，年龄和购买总金额对于客户流失的影响并不大。

有客户经理跟进的客户反而更容易流失，服务时间越长的客户也更容易流失，可以看出使用公司服务的客户网站数量对于流失率影响较大，数量越多的客户越容易流失。通过观察后客户经理、服务年数和使用公司服务的客户网站数量可以得到客户流失的原因，进一步进行决策。



1. 特征重要图

特征重要图反映了不同特征对于客户流失的影响，其可视化结果如下图所示。可以看出，公司服务的客户网站数量以及总消费是最重要的特征，后续可以根据重点关注具有此类特征的客户流失预防。



## 4 小结

本实验通过选取用户流失预测作为主题，结合数据挖掘、机器学习的相关方法，使用PySpark用户流失预测。实现本案例的主要内容有：（1）采用逻辑回归和梯度提升树方法进行用户流失的分类预测，并使用AUC指标和PR指标进行评估。（2）使用K折交叉验证进行模型超参调优。（3）绘制特征与流失用户关系图、特征图重要度实现结果可视化。

## 附录

1. 本案例提供配套的PPT、视频、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例涉及到数据处理以及机器学习算法，并使用PySpark进行实现。

3. 本案例参考文献如下：

[1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

参考内容：逻辑回顾、梯度提升树、超参调优、交叉验证法、AUC指标、PR指标