Spark 项目 2020年3月11日
数据挖掘直通车班

预测Spotify平台用户流失情况

项目概述

Spotify平台采用订阅模式，盈利主要来源于广告以及订阅费用。为了提高平台利润，需要尽可能多地吸引用户并且尽力避免用户流失。为了尽可能挽留即将流失的用户，需要给予一定的优惠服务。平台收集用户的使用数据，根据用户的注册信息以及过往的行为历史，使用机器学习方法，预测用户流失的概率。

问题陈述

项目目标是预测用户流失的概率，需要完成以下步骤：

1 加载和清洗Spotify平台数据
2 探索性数据分析，充分了解数据集
3 特征工程，筛选/建立与预测用户流失情况相关的重要特征
4 使用特征工程建立的特征，搭建模型预测用户未来流失的情况
5 根据评估指标进行模型优化

该分析建模过程暂时使用完整数据集的子集进行

评估指标

预测用户流失情况属于监督学习中的分类问题，由于流失用户人数与留存用户人数之比大约为1:3，分布不平衡，所以使用F1 Score，F1 Score兼顾了precision（误报率）与recall（检出率），是precision与recall的几何平均数。

加载和清洗Spotify平台数据

排除userId为空的数据，因为这些数据是在用户未登录的状态下产生的，与本项目预测用户流失情况的需求无关

探索性数据分析 & 特征工程

1 对数据集中流失用户与留存用户进行标记，发现在这个小数据集中流失用户与留存用户的比例约为1:3。
2 数据产生于2018年10月1日至12月31日。
3 流失用户平均活跃天数为23天，留存用户平均活跃天数46天

4 用户性别分布大致均衡，有部分用户同时拥有free与paid两个level

由于我们需要做的是根据用户的注册信息以及历史行为（浏览网页的时间分布）预测未来用户流失还是留存。在探索性分析中要通过分析找到流失用户组与留存用户组两组之间存在差别的特征。 要排除两组之间的人数不同以及平均活跃时间不同这两个因素对用户行为统计结果的影响。

因为后续使用机器学习方法的目的是预测用户流失的概率，以对流失概率高的用户采取一定的优惠措施。机器学习方法针对的是个体而不是总体，所以两组之间的样本数量差异会干扰机器学习方法的特征选取。在预测的时候某名用户在未来留存或流失是未知的，所以样本每名用户的活跃时间对机器学习方法没有参考价值

基于控制变量的考虑排除两组之间人数的差别以及平均活跃时间不同这两个因素可以更好地考察两组之间其他特征的差异（如播放音乐的数量）

经过探索性分析，有以下特征可以用来预测未来用户留存或流失：
1 用户每天平均Roll Advert的数量
2 性别
3 用户平均每天Thumbs Down数量
4 用户平均每天Downgrade数量
5 用户每天平均播放音乐数量
6 用户平均每天Add to Playlist数量
7 用户平均每天Help数量
8 用户平均每天Setings数量

建模 & 根据指标进行模型优化

由于Spark本身的局限性，机器学习算法选用随输入数据大小线性扩展到算法，项目处理的是分类问题，可以选用LogisticRegression, LinearSVC, RandomForest, GradientBoosting四种分类器。

选择F1 Score作为评估指标，运算时间也是重要的指标。将数据分为训练集与测试集，在训练集上使用交叉验证方法，网格化搜索最优的参数组合，对比每种模型在这个子数据集上的最优解，发现综合考虑上述评估指标，RandomForest分类器的在训练集上表现最优的参数组合在测试集的f1-score达到了0.826。

改进

1 模型还有进一步改进的空间，如果使用更精确的GridSearch，但代价是需要更长的模型训练时间和成本。另外Spark能够使用的机器学习方法远少于sklearn等机器学习库，还有一些机器学习方法未能使用，比如神经网络/朴素贝叶斯等，神经网络可能带来更好的效果，但

在大数据集上训练的成本也更高。如果Spark支持RandomSearch，BayesianSearch，可能做到更有效地搜索参数组合。

2 这只是在Workspace的子数据进行的分析与预测，下一步需要在AWS的完整数据集上进一步分析与预测。有一些用来预测用户流失情况的特征可能不适用于完整数据集，整个流程有可能要重新调整。