Sparkify Project YiSongZhang

数据挖掘工程师纳米学位 2/24/20

定义

项目概括

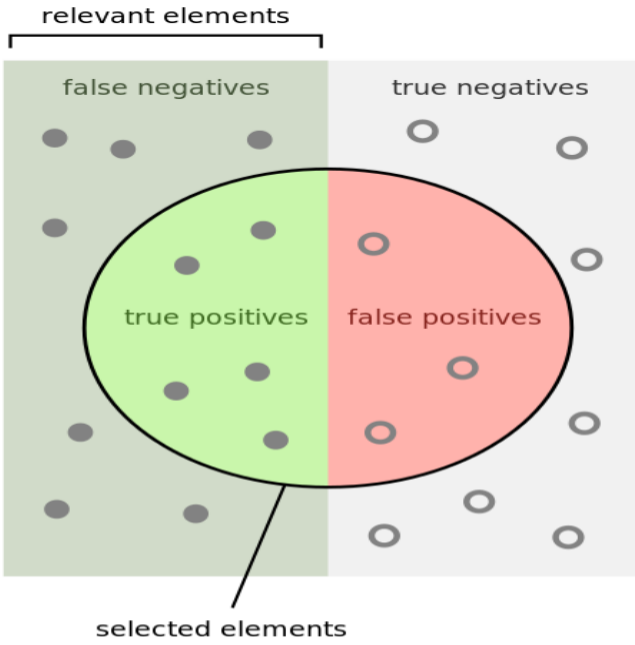
在音乐公司 Sparkify，用户流失是一个关键问题，用户量跟利润有很大的相关性。因此准确预测用户流失是极为重要的，找到和用户流失的关键，然后改进产品从而降低用户的流失。在这个项目中，我根据 Udacity 提供的数据，经过分析和建模来预测用户的流失情况。

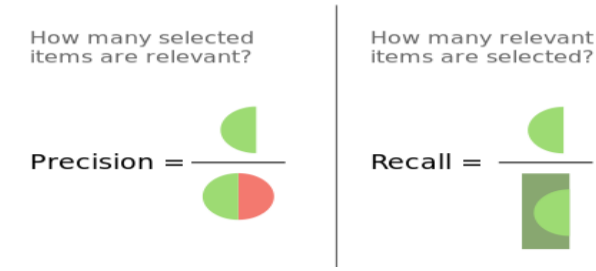
问题陈述

需要利用 spark 来处理数据，预测用户流失是一个二元分类问题，因此要用机器学习分类算法来进行预测，流程包括数据清理，探索，特征工程，建立模型，模型评估。

指标

F1 分数是用来衡量二分类模型精确度的指标，它同时兼顾了模型的精确率和召回率，精确率和召回率在这里都对我们的预测结果很重要，因为所有流失用户都希望被准确预测，并且减少将留存用户预测为流失用户。F1 分数的最大值是 1，最小值是 0。





$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

分析

数据清洗

数据集大小为 128MB，为完整数据 12GB 的子集。

userId 列有空字符串的情况，把相关行全部删除。

location 列有整洁度问题，应该转换成地址和地区两列，因为不需要这一列特征，在此不做处理。

数据探索

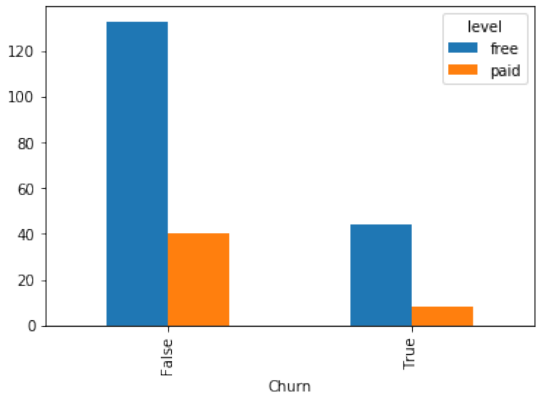
原始数据结构一共有 286500 行，18 列，具体如下所示：

```
root
 |-- artist: string (nullable = true)
 |-- auth: string (nullable = true)
 |-- firstName: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- itemInSession: long (nullable = true)
 |-- lastName: string (nullable = true)
 |-- length: double (nullable = true)
 |-- level: string (nullable = true)
 |-- location: string (nullable = true)
 |-- method: string (nullable = true)
 |-- page: string (nullable = true)
 |-- registration: long (nullable = true)
 |-- sessionId: long (nullable = true)
 |-- song: string (nullable = true)
 |-- status: long (nullable = true)
 |-- ts: long (nullable = true)
 |-- userAgent: string (nullable = true)
 |-- userId: string (nullable = true)
```

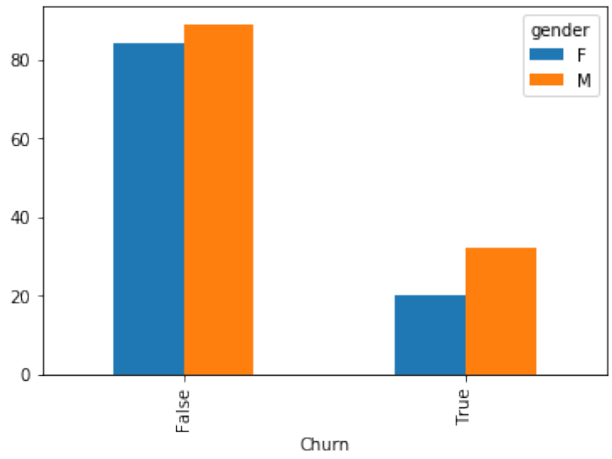
通过 page 页面是否为 Cancellation Confirmation，我发现数据集里一共有 52 个用户注销了账号。并且我给数据集新加了一列，名为 Churn，如果值为 1 就代表这一行的相关用户为注销用户，0 为非注销用户。

数据可视化探索

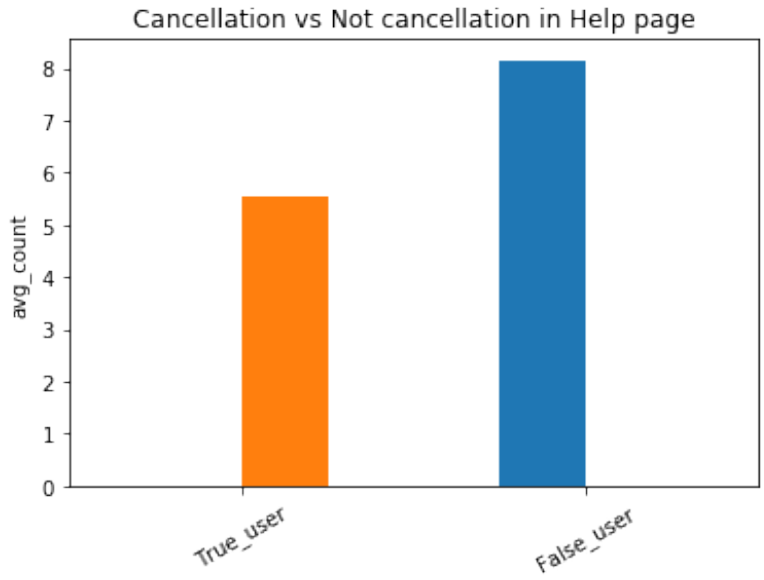
- 可视化在注销用户与非注销用户里面的用户类型数量对比，可以看到在注销用户里付费用户相对少一些，level 可以放在特征里。



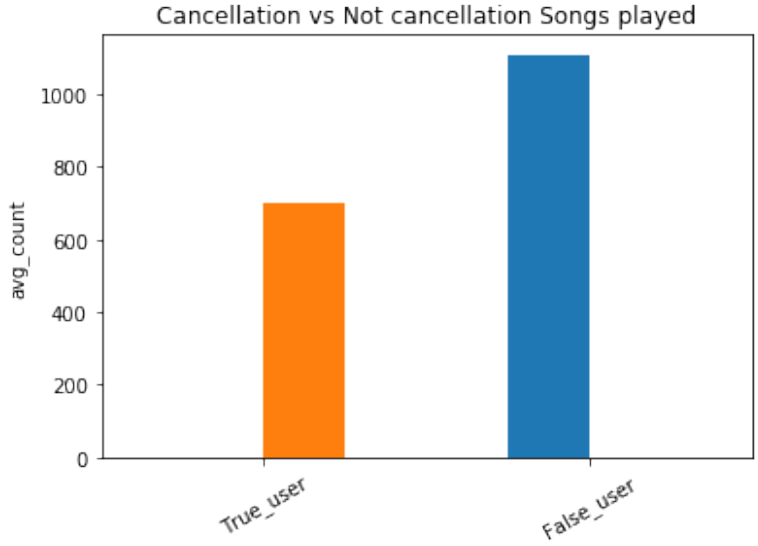
- 可视化在注销用户与非注销用户里面的性别数量对比，可以看到，注销用户里男性要相对多一些，性别可以放到特征里。



- 分别找出注销用户和非注销用户平均每个人的帮助页面点击量，然后合并到一张数据框上，用柱形图对比注销用户和非注销用户的人均帮助页面点击量，`True_user` 为注销用户，`False_user` 为非注销用户，可以看到，注销用户人均点击帮助页面量要比非注销用户低，我想可能是流失用户遇到软件使用困难的时候软件没有给到有效的帮助指引，从而导致用户流失，所以我觉得帮助页面点击量可以放在特征里。

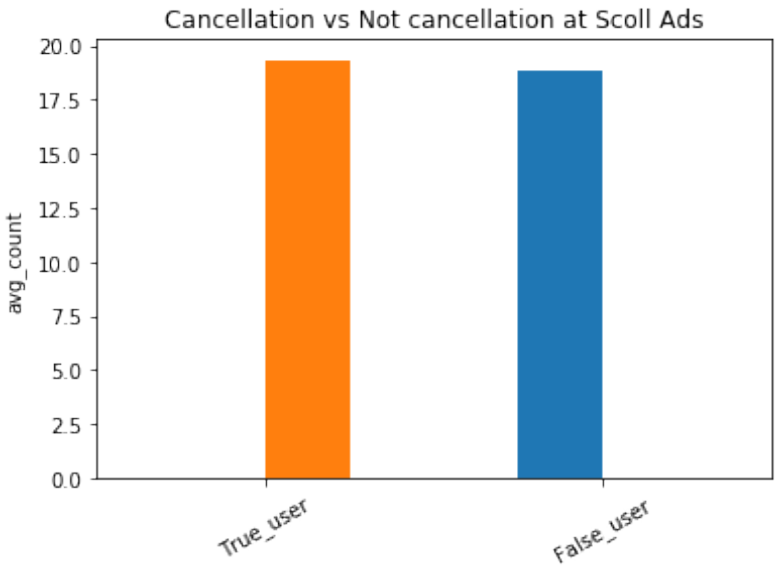


- 用柱形图对比注销用户和非注销用户的人均听歌量，`True_user` 为注销用户，`False_user` 为非注销用户。可以看出注销用户人均听歌量要比非注销用户少很多，或许是因为推荐系统给注销用户推荐的歌不匹配，也有可能是我们曲库里没有他们喜欢听的歌的版权，所以听歌量可以放在特征里。



- 用柱形图对比注销用户和非注销用户的人均点击广告量，`True_user` 为注销用户，`False_user` 为非注销用户，可以看出注销用户人均点击广告量

要比非注销用户高一些，我想太多的广告体验或许会影响用户的流失，所以我认为广告点击量可以放在特征里。



方法与结果

特征工程

最终我选择了 `Ads`，`Songs`，`Help`，`level`，`gender` 特征来进行特征工程，`Churn` 作为标签使用。

模型选择

我选择了三种模型：

LogisticRegression：（F1 Score：0.7840683558206027）

GBTClassifier：（F1 Score：0.8040928006815921）

RandomForestClassifier：（F1 Score：0.9342054807211727）

效果最好的模型是 RandomForestClassifier，得到的 F1 分数约为 0.93。

关于随机森林模型，我第一次添加了太多参数分别是 numTrees, minInstancesPerNode，maxDepth，minInfoGain，并且参数量很大，导致模型运行的太慢了。然后我取消了一个 minInfoGain 参数，但是第二次的结果标签全部为 0，因此我把参数名字调整的和第一次一样，但是把 minInfoGain 调整小分别为 0 和 1，minInstancesPerNode 分别调整小为 1 和 3，我把 numTrees 和 maxDepth 分别调小来让模型运行速度提升，结果还不错。 最好的参数为： numTrees:10 minInstancesPerNode:1 maxDepth:7 minInfoGain:0.0

总结

结果讨论

本项目中的问题是预测流失用户，也就是一个二元分类预测问题，所以所选择的模型要契合这一特点效果会比较好，我用了随机森林分类器，GBT 分类器和逻辑回归分类器，效果最好的是随机森林分类器，F1 分数达到了非常高的 0.93，首先我对数据集进行了清洗，然后用 EDA 方法对数据集进行了探索，找出特征之间的规律，对需要的特征进行特征工程，最后建立机器学习模型对其预测。随机森林分类器专门用来解决分类和回归问题，而且可以同时处理分类和数值特征并且抗过拟合能力较强，我想这也是在训练集和测试集都表现良好的体现。但是由于项目提供的数据集较少，因为网络问题无法使用 IBMCloud 等处理大数据强大工具，在本机上运行模型收到硬件和网络的限制，调整参数变得更加困难，如果这些方面可以提升，我觉得结果会更好。

困难与挑战

在最后建立模型的时候有很多模型可以用，但是经过我测试了很多模型发现，有的模型会出现难以解决的错误，比如说我在使用 NaiveBayes 模型的时候就出现了复杂的错误，在网上很难找到解决方案，与软件版本号有关，这时候就要换一个模型。还有在调整参数的时候刚开始输入了不合适的参数导致效果非常慢，后来在网上对每个参数进行搜索和深度了解后优化参数，这一步比较难，因为牵扯到时间成本。还有在可视化的时候我需要把一个 pandas 数据框的列转换成行然后直接转化成我要的柱状图，我试了一些方法都没有成功，直到在网上查到了 melt 和 pivot 混合使用方法才解决。

有趣的发现

在我用柱形图对比注销用户和非注销用户的人均帮助页面点击量时我发现了有意思的事情，结果显示注销用户的人均帮助点击量要比非注销用户的少很多，我的主观意识一开始告诉我，点击帮助页面越少应该代表遇到使用问题越少导致软件使用体验越好才对，那为什么流失用户点击量要比留存用户少呢？我仔细想了想，可能是流失用户遇到软件使用困难的时候软件没有给到有效的帮助指引，或者是用户看不懂，也有可能是用户不知道去哪里找帮助页面造成的，而不是真的没有遇到使用问题。

反思与改进

在进行特征工程的时候我把 gender 和 level 两个特征分别转化为 0 和 1，但是我忘记把它们转化为整数类型了，从而导致了在建立模型的时候有类型错误，所以在处理特征的时候一定要特别注意数据类型的转换。在建模的时候我选择了很多模型，其中有几个因为出现了比较难解决的错误，包括牵扯到一些软件

版本问题，我只能重新筛选模型，对比，找到最适合的几个。由于我没有使用IBM Cloud 和 AWS，所以在调参的时候刚开始经历了很多次运行漫长的经历，后来通过调整合适的参数，减少参数来解决这一问题，在运行模型的时候不光要考虑效果也要考虑效率，综合硬件和环境能来用最快的时间建立最优的模型。

参考

https://cn.udacity.com/

http://spark.apache.org/docs/2.0.0/api/python/_modules/pyspark.html

https://pandas.pydata.org/

感谢

Udacity

Mentor:Ray

Classmates:Allen