Spyrkify Project

Big Data Nanodegree

Huang Chao

November 28, 2019

定义 Definition

项目概览 Project Overview

用户流失是每个软件公司不愿意看到的事情，对于音乐公司也是如此。因此，根据用户以前的数据，预测可能的流失用户，在用户注销之前提供各种可能的挽留措施，能够最大限度减少软件公司的用户损失情况。

在这篇报告中，我根据 Udacity 提供的数据，在 IBM Watson Studio(free version) 云平台上运用 Pyspark 完成了对该数据集的大数据处理分析预测。

问题陈述 Problem Statement

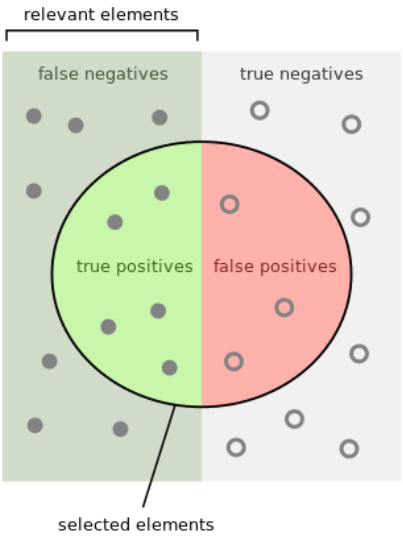
预测客户流失率是数据科学家和分析师在面向消费者的一类公司中经常遇到的一项具有挑战性的问题，能用 Spark 高效处理大数据集是数据领域职位急需的一种能力，主要挑战包括:

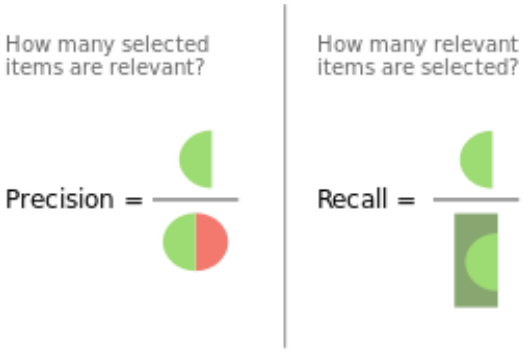
1. 把大数据集加载到 Spark 上，并使用 Spark SQL 和 Spark 数据框来操作数据。
2. 在 Spark ML 中使用机器学习 API 来搭建和调整模型。

预测流失用户，本质上是针对用户的二分类问题，因此适合采用监督学习的机器学习方法。项目主要流程包括：数据清洗，探索性分析（EDA），特征工程，监督学习建模，建模过程中采用 f1 分数对模型进行评估和改进。

衡量指标 Metrics

我们希望所有流失的用户能够被准确的查到，减少第一类错误发生的概率，因此查全率（recall）是衡量模型的一个重要指标，同时，我们也希望尽可能少的将留存的用户预测为流失用户，即减少第二类错误的概率，因而准确率（accuracy）指标也要考虑。所以，在项目中 F1_score 成为我们评价指标的首选。





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

分析 Analysis

数据清洗 Data Wrangle

完整的数据集大小为 12GB，由于计算条件的限制，我们选取了它的一个迷你子集，大小约为几百兆，并在免费的 IBM 云完成大数据任务的处理。

质量：

1. ‘uerId’一列中存在空白的记录：删除了没有 userId 记录的行

清洁度：

1. 拆分 location 一列，转换为 address 和 area 两个特征

数据探索 Data Exploration

数据主要形式如下：

```
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)
```

定义客户流失

在完成初步分析之后，使用 Cancellation Confirmation 事件来定义客户流失，创建一列 Churn 作为模型的标签，该事件在付费或免费客户身上都有发生。

探索可视化 Exploratory Visualization

Fig. 1 迷你数据集中，注销的用户有 99 个，未注销的用户有 349 个，注销的用户比例约为 22%，还是比较高的，因此预测这一部分用户是很有意义的。

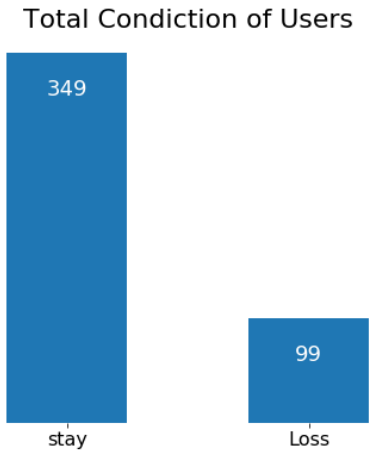


Fig. 2 下图表示用户听歌时间的分布图，黄色表示没有注销的用户，蓝色表示注销的用户，可以看到，两者的趋势都是相似的，在傍晚这段时间听歌的人最多，但是注销的用户趋势没有其他的明显。

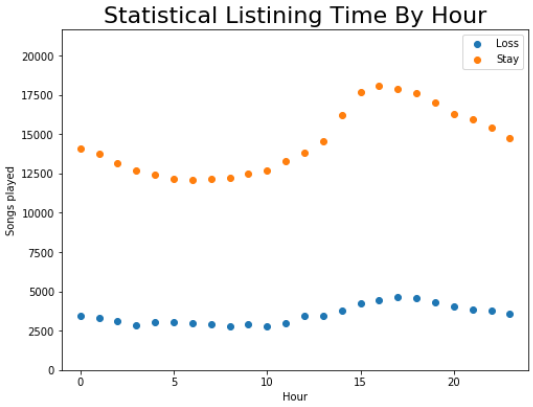
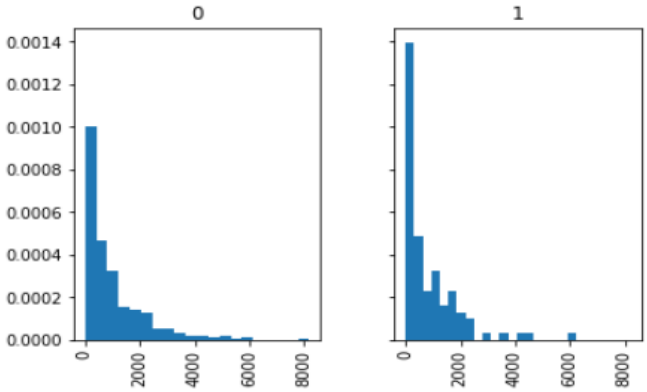


Fig. 3 下图表示用户的听歌总数，其中 1 表示注销的用户，可以看到，注销的用户显然听歌的数量较少，从中位数来看，注销的用户仅为 439，而留下来的用户高达 601，因此这个特征可以纳入特征工程。



其他变量也遵循类似的步骤，我们选取了 thumbsdown，thumbsup，rolladvert，addtoplaylist，home 等特征作为后续分析的一部分。

方法与结果 Methodology And Result

特征工程：

基于以上的分析，最终初步选取了 churn，gender，songs，home，thumbs up，thumbs down，roll advert，add to playlist，artist，level，add friend 这些变量作为我们的特征工程。其中 churn 作为标签使用。

模型选择：

尝试了三种模型：LogisticRegression，RandomForestClassifier，GBTClassifier，其中 LogisticRegression 模型效果最好，在测试集上 f1_score 为 0.76，因此选择该模型进行预测。

- LogisticRegression (best f1 score 0.7714)
- GBTClassifier (best f1 score 0.7214)
- RandomForestClassifier (best f1 score 0.7276)

结果与讨论：

由于迷你数据集的数据还是太少，因此预测的准确率三个方法都偏低。在预测测试集上的数据集时，一开始选择了 AUC 指标（对于二分类问题，这个指标能够较好的反馈我需要的结果，并且 Pyspark 中的 BinaryClassificationEvaluator 提供这个评估指标）。但是在对测试集预测的详细检查中发现，0.66 的 AUC 值，预测的结果全为 0，这个是不愿意看到的结果。这时，不得已选择了 MulticlassClassificationEvaluator 的 f1 指标，并对逻辑回归的阈值作了一定的调整（threshold=0.4），得到了较好的结果，f1 指标为 0.76。对于不加权的 f1 指标，也有 0.36，虽然比较低，但是由于数据集实在是太少，这个预测结果已经令人满意了。

总结 Conclusion

预测流失用户，本质上是针对用户的二分类问题，因此适合采用监督学习的机器学习方法。项目主要流程包括：数据清洗，探索性分析（EDA），特征工程，监督学习建模，建模过程中采用 f1 分数对模型进行评估和改进。

在这个问题上，由于特征较少，所以随机森林算法没有优势，逻辑回归算法表现更好。

对于项目改进，可以提取更多的特征，增加数据量，以及调试更多的参数改进模型。进一步的计划：对于该模型，将用户随机分成基于账号 ID 的两组，一组对预测为流失的用户进行优惠活动，一组照常，然后以顾客流失率作为指标进行 A/B 测试。

困难与挑战

- 一开始想在原始数据上进行建模，但是非常的麻烦，采用了独热编码等很多方法，后来发现独立出来后 join 比较方便。
- Pyspark 的预处理方式很有趣，不过画图的时候需要转换为 pandas，稍微有点麻烦。
- 在选择评价指标的时候出现了一些问题，这个使我意识到，对于一个模型，评价指标非常的重要，要谨慎选择。

参考：

https://github.com/Mikemraz/Capstone-Project-Big-Data-Sparkify

感谢：

udacity 提供课程和相关的数据