|  |  |
| --- | --- |
| **Spyrkify Project** | **Huang Chao** |
| Big Data Nanodegree | **November 28, 2019** |

**定义Definition**

**项目概览**

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。还有，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：

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。
3. 将你在 Spark 课程和数据科学家毕业项目中学到的技能结合起来。

**衡量指标 Metrics**

F1\_score：综合了查全率和准确率，能够较好的评价模型的效果。

**分析Analysis**

**数据清洗Data Wrangle**

完整的数据集大小为 12GB，由于计算调价你的限制，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。Udacity提供的数据集相对来说比较整洁，因此数据清洗工作量较小。

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：

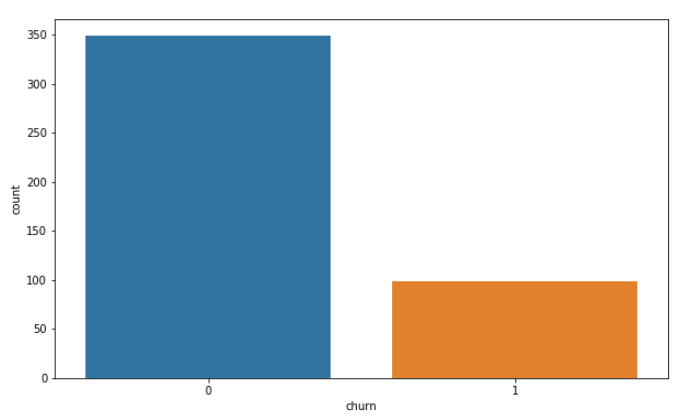
1. 暂无

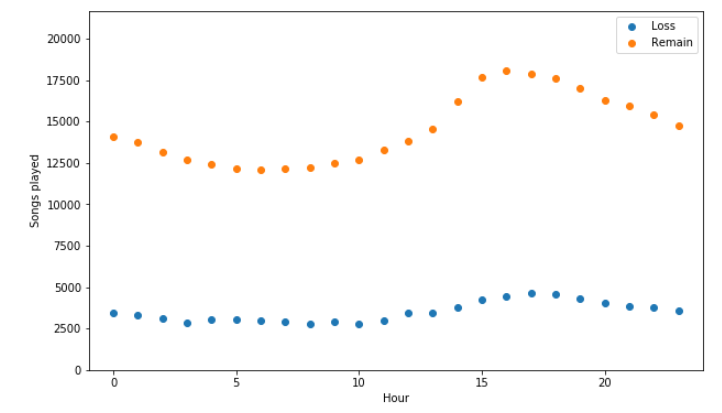
**数据探索 Data Exploration**

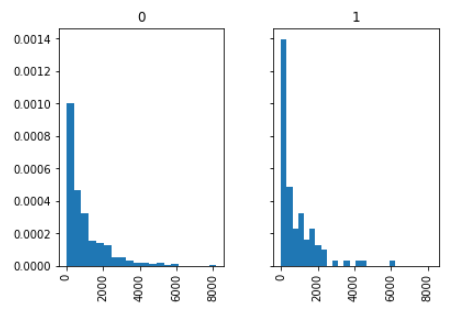
**定义客户流失**

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0.74，效果令人满意。

参考：

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

感谢：

udacity提供课程和相关的数据