# QG 工作室暑期实习生两日结

姓名： 许级元　　组别： 数据挖掘　年级： 19 级　　日期： 2020.08.08

| | |
|---|---|
| 生活小记 | 8.7 日，有一个朋友也是学大数据方向的，来找我一起打阿里天池的新人实战赛，于是就开始了准备，由于可以使用框架，也想趁机了解一下 XGBoost 框架的基本使用。数据集是 o2o 优惠券使用预测，数据内容为用户在线下以及线上的领券以及消费行为记录。上午下载完数据，我们对数据进行了预处理以及做了特征工程，数据预处理具体为时间戳转换、距离处理以及折扣率的处理，划分完数据集之后再从训练集、验证集和测试集上划分历史区间、中间区间以及标签区间。然后对数据集打上标签。特征工程方面，首先从已有数据集上提取了用户领券数等特征，接着根据用户领券日期提取日期特征，最后再对数据进行抽样训练，筛选出相关度高的特征。下午基于上午做的工作，构造出新的数据集，然后调用 XGBoost 库来构建一个 XGBoost 模型，调整好参数之后训练了几轮，最后确定好了模型参数，整理了一下代码文件。晚上接着下午的工作，按比赛要求整理好结果文件的输出，最后在阿里云上提交了结果文件，AUC 是 0.72 左右，而在训练集是 0.8 左右，显然模型过拟合了，可能是第一次使用 XGBoost 又或者是特征工程方面还是有欠缺，不过总算是完成了这次比赛。晚上睡觉前依旧抽出时间练练琴。<br>8.8 日，上午学习了 HMM 这一著名的有向图模型，之前就有看过 HMM 的介绍，学起来还是蛮抽象的。下午了解了 MRF 和 CRF 这两个无向图模型，不过还没有实现其代码。晚上是设计组的技术交流会，了解到了 AE 和 PR 的区别，前者偏向于做视频特效，后者偏向于做视频剪辑。技术交流会结束后就是康乐活动了，还是和上次一样，和小伙伴以及师兄打王者荣耀，不过这次没有上头，打了几盘之后就溜了。 |
| 学习<br>开发<br>比赛<br>概要 | 学习上，学习了 HMM、MRF 以及 CRF 这三个模型。另外学习了 XGBoost 这一机器学习框架。<br>比赛上，完成了阿里天池新人实战赛，第一次尝试使用 XGBoost 框架，尝试了处理一些与时间相关的数据特征。 |

| | |
|---|---|
| **感想收获** | 平时很少时间去尝试数据挖掘比赛，这次朋友的邀请是一个难得的机会，可以学习到新的机器学习框架以及积累一些数据预处理和特征工程的经验。感觉收获还是蛮大的。<br>康乐活动是放松的好机会，劳逸结合才能带来好的学习效率。 |
| **存在问题（备注）** | |