项目：音乐网站用户流失预测

项目成员：陈湘孔、马亚军、任崇伟（姓氏顺序）

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

刚开始讨论：

思路：

经过讨论，本项目初步分为四步，包括数据探索、特征工程、模型训练、测试集上预测性能。

其中模型训练分为两步：

1. 协同过滤
2. CTR预估排序

工作要求：

* 数据探索

写出数据探索文档，包括对数据集的分析，总结

* 特征工程

协同过滤阶段：对缺失值处理；离群点处理

CTR预估阶段：特征融合；

* 模型训练

再定…

* 性能预测

再定…

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

后来的思路：

1. 模型训练阶段，得到user-item-score矩阵，得到user-item、item-user、(bi、bu、qi、pu这四个矩阵很关键)
2. 测试阶段：

* 数据预处理：去掉冷启动用户
* 输入用户id，对于每一个用户id，都会返回一个分数，那么user-rating就是要提交的结果