性别模型

业务大背景如下：

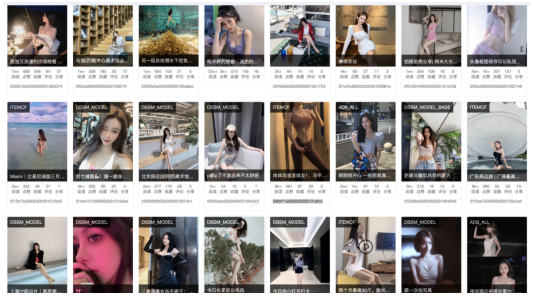
平台推荐上会对男性，甚至女性用户出现局部最优的，过拟合的内容推荐。影响了时间维度上的全局最优，需要根据大家的普遍认识，来识别用户的真实性别。

美女内容处理

背景

https://david.devops.xiaohongshu.com/#/history/572c516e5e87e7305ef4006a/homefeed

- 对男性用户，美女价值压倒了其他内容价值，且美女价值是跨品类的，基于类目的打散机制失效
- 不利于男性用户发现其他更长期的产品价值
- 打压的美女价值，而不是美女笔记，调整现有大尺度的标准胖大于利，全标代价大（去除暴露的，还有穿着端庄长得好看的，总有相对漂亮的那一波）



识别策略

识别 → 打散（处理）
https://gravity.devops.xiaohongshu.com/api/gu/s/ad44fbc8-d3d9-11ea-8983-0a58ac16bf7a

CV @张一帆

- 标准：女性脸、身体部位为图片主体（可选：颜值高身材好皮肤好等）（focus在封面）
- 数据集准备
  - 优选：先识别人体，初筛结果作为待标注数据集
  - 次选：基于用户行为，从男女行为diff找一些笔记初筛作为待标注数据集
  - 兜底：挑部分类目，抽样作为数据集
  - 8.3讨论数据集准备方案：@张无忌 (何永能)
    - 用现有的种子笔记，通过相似笔记扩展更多数据集
    - 种子笔记（100篇）：美女笔记种子样本.xlsx
    - 利用相似向量，拓展1K、5K、1W篇，看下准确率
  - 次选：基于用户行为，从男女行为diff找一些笔记初筛作为待标注数据集
  - 兜底：挑部分类目，抽样作为数据集

行为侧 @数据分析师 @非白、无忌

- 统计侧：note breakdown，人均click（dau base，即click/dau），男性显著高
  - 大盘男女情况如下
viewergender	COUNT_DISTINCT(userid) 占比%	SUM(click) 占比%
–	–	–
1	80.42%	85.96%
0	14.88%	8.94%
2	4.70%	5.10%
- 模型侧：在笔记侧用行为相似性做一些尝试

【待定】真男人识别 @吹雪

- 从行为特征上计算性别（可能鸡生蛋蛋生鸡，先做了内容识别，更好甄别用户性别）
- 历史文档
  - 非实时数据数据质量摸底
  - https://code.devops.xiaohongshu.com/data/RED-Deeplearning/tree/release/userprofile

实验逻辑（处理策略）

控制组：线上逻辑
实验组1： 对男性，满足CV的笔记，打散n出1
实验组2： 对男性，同时满足CV + 行为特征的笔记，打散n出1

模型主要建立的假设：

1. 在已经填写了性别的用户当中，存在真的男，假的男，真的女，假的女四类
2. 男性和女性（内心）在行为上存在差异，不同性别用户行为（曝光、点击、关注、收藏、发布等）上具有一定倾向性
3. 用户存在多种喜好倾向
4. 依据倾向下手填性别的比例和数量，可以推断这种倾向属于男性还是女性，及可信程度
5. 依据用户多种倾向的性别归属及可信度，可推断用户的真实性别

数据：

1. 笔记taxonomy
2. 内容发布
3. 用户→ 笔记
    a. click
    b. 赞藏
    c. 时长
    d. 评论
4. 用户→targetUser
    a. 笔记
    b. 私信
    c. 性别
5. 用户昵称
6. 手填性别