字节跳动 安全中心 × 安全范化 BYTEDANCE SECURITY

色情导流用户识别

-ECNU_ICA

im0qianqian、Bin、RR

字节跳动
安全AI挑战赛

字节跳动 安全中心 × 安全范化 BYTEDANCE SECURITY

目录

字节跳动
安全AI挑战赛

Empower Security
Enrich life

1 团队简介

字节跳动
安全AI挑战赛

字节跳动 安全中心 × 安全范化 BYTEDANCE SECURITY

团 队 简 介

Bin

im0qianqian

RR

字节跳动
安全AI挑战赛

Empower Security
Enrich life

2 赛 题 理 解

字节跳动
安全AI挑战赛

字节跳动 安全中心 × 安全范化 BYTEDANCE SECURITY

任务简介

字节跳动
安全AI挑战赛

Empower Security
Enrich life

评估指标

根据实际风控业务中，色情导流用户的数量与正常用户相比有较大差距，如果准确率较低会召回大量正常用户，所以采用 F-beta 作为模型的评估值（beta=0.3），该值**越大越好**

$$F-beta=(1+beta^2)\frac{Precision*Recall}{beta^2*Precision+Reall}$$

3 特 征 工 程

字节跳动
安全AI挑战赛

字节跳动 安全中心 × 安全范化 BYTEDANCE SECURITY

基础特征

字节跳动 安全中心 × 安全范化 BYTEDANCE SECURITY

用户基础特征

- gender_str，性别，离散型
- signature，个性签名，序列型
- create_time，账户创建时间，连续型
- follow_num_all，关注人数，连续型
- fans_num_all，粉丝数，连续型
- publish_cnt_all，投稿数，连续型
- server_comment_cnt_all，评论数，连续型

用户行为特征

- video_play，播放次数，连续型
- video_play_finish，播放完成次数，连续型
- play_time，播放时长，连续型
- click_video_play，点击视频播放次数，连续型
- feed_request，feed 请求数，连续性 **（无用特征）**
- homepage_hot_slide_up，feed 页上滑数，连续型
- homepage_hot_slide_down，feed 页下滑数，连续型
- like，点赞数，连续型
- dislike，点不喜欢数，连续型
- post_comment，评论数，连续型
- search，搜索数，连续型
- share_video，分享数，连续型

用户投稿特征

- item_title，投稿标题，序列型
- poi_name，投稿所在 POI，离散型
- item_province_cn，投稿所在省份，离散型
- item_create_time，视频投稿时间，连续性

色情导流用户 导流方式调查

二维码

视频中或者头像上出现的二维码

联系方式

在个签，视频，视频描述中出现的常用通讯工具的联系方式

短链接

通过短链接诱导用户去点击进入网站

诱导性话语

通过谐音字/象形字达成导流方式的建立

多重用户指向

在一个账号的视频下通过@，评论，转发等方式导流用户进入到下一个账号

同城功能

利用时间段人性弱点，在特定时间段内进行相关信息的投放

字节跳动
安全AI挑战赛

Empower Security
Enrich life

领域特征挖掘与构建 – 调研研究

个人主页中的引流方式：

- 使用 emoji 代替文字内容
  - 例如『v💗、同🐂』
- 使用同音字代替文字内容
  - 例如『加薇、+薇、+v』
- 带有分隔联系方式的冒号
  - 例如『VX:、QQ:』
- 头像中包含引流二维码

也存在正常的商业导流，其行为差别在于**是否刻意使用语义技巧隐藏信息**

字节跳动
安全AI挑战赛

Empower Security
Enrich life

领域特征挖掘与构建 – 调研研究

视频 POI 和描述高度相符，通过同城引流

- 个签/账户名中包含城市名
- “同城”中往往深夜时段色导用户较为活跃

字节跳动
安全AI挑战赛

Empower Security
Enrich life

领域特征挖掘与构建 – 特征构建

字节跳动 安全中心 × 安全范化 BYTEDANCE SECURITY

个签及投稿标题的序列特征

- 色导用户常使用谐音字/象形字/emoji替代一些文字，正常用户则不会
- 导致色导用户个签中存在标志性字符

对文本序列信息建模：

- 拼接用户个签及所有投稿视频标题
- 基于 TF-IDF 抽取前 3 个关键字做特征
- 基于词袋模型编码文本并作为特征

领域特征挖掘与构建 – 特征构建

字节跳动 安全中心 × 安全范化 BYTEDANCE SECURITY

明文链接特征

- 色导用户个签及投稿视频标题中或许存在链接进行导流
- 因此考虑基于规则鉴别链接信息（http:// 与 https://）

字符	index
h	2189
t	7579
p	4010
s	1385
:	1712
/	7571

领域特征挖掘与构建 – 特征构建

字节跳动 安全中心 × 安全范化 BYTEDANCE SECURITY

个签中字母 / 数字 / 符号数量

字符	index
h	2189
t	7579
p	4010
s	1385
:	1712
/	7571

+

- 提取 URL 并分析
- 迭代获得 URL 中存在的字符、数字以及部分符号所对应的 index

→

- 统计个签字母 / 数字 / 符号的数量并作为特征

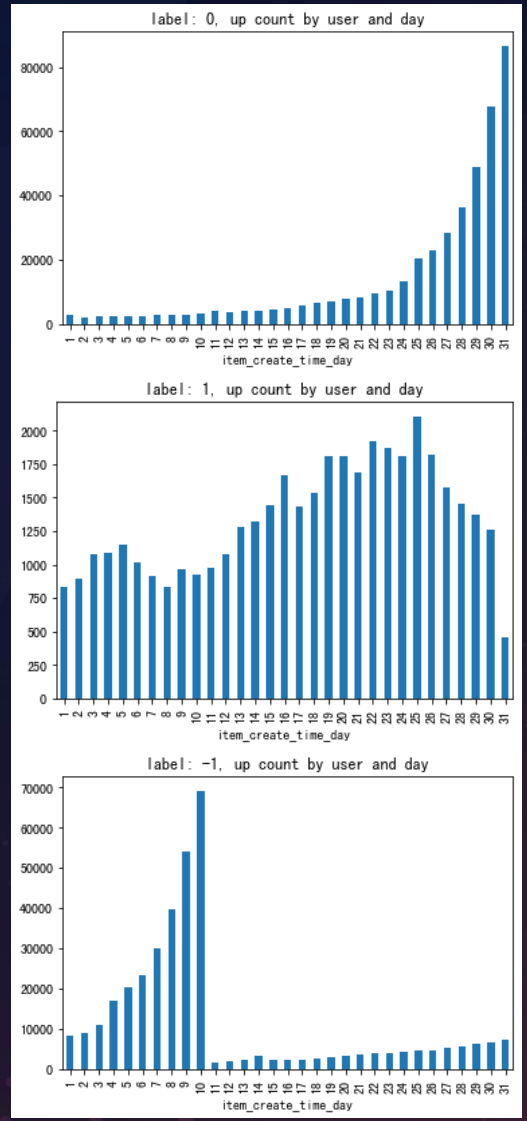
特征聚合

考虑聚合操作有助于挖掘额外的统计特征，例如：

- 每一个用户的投稿数量
- 每个用户投稿地区覆盖多少个省份 / POI
- 个性签名为空的用户平均粉丝数是多少

需要注意的一点：

- 用户投稿时间在训练集及测试集分布差异巨大（包括时间跨度不一致）
- 因此对于投稿时间仅保留时间大小无关的特征（例如平均投稿时间、总投稿时间跨度、投稿时间方差标准差等）



训练集

测试集

用户投稿时间（day）

字节跳动
安全AI挑战赛

Empower Security
Enrich life

特征交叉

考虑特征之间的交互容易组合出一些更为有用的新特征，例如：

- 播放时长 / 播放次数 = 平均播放时长
- 粉丝数 / 投稿数 = 平均每个投稿所获粉丝量

交叉方式主要包含：

- A * B
- A + B
- A – B
- A / (B + 1e-7)

4 模 型 构 建

字节跳动
安全AI挑战赛

字节跳动 安全中心 × 安全范化 BYTEDANCE SECURITY

机器学习模型

深度学习模型

阈值选择

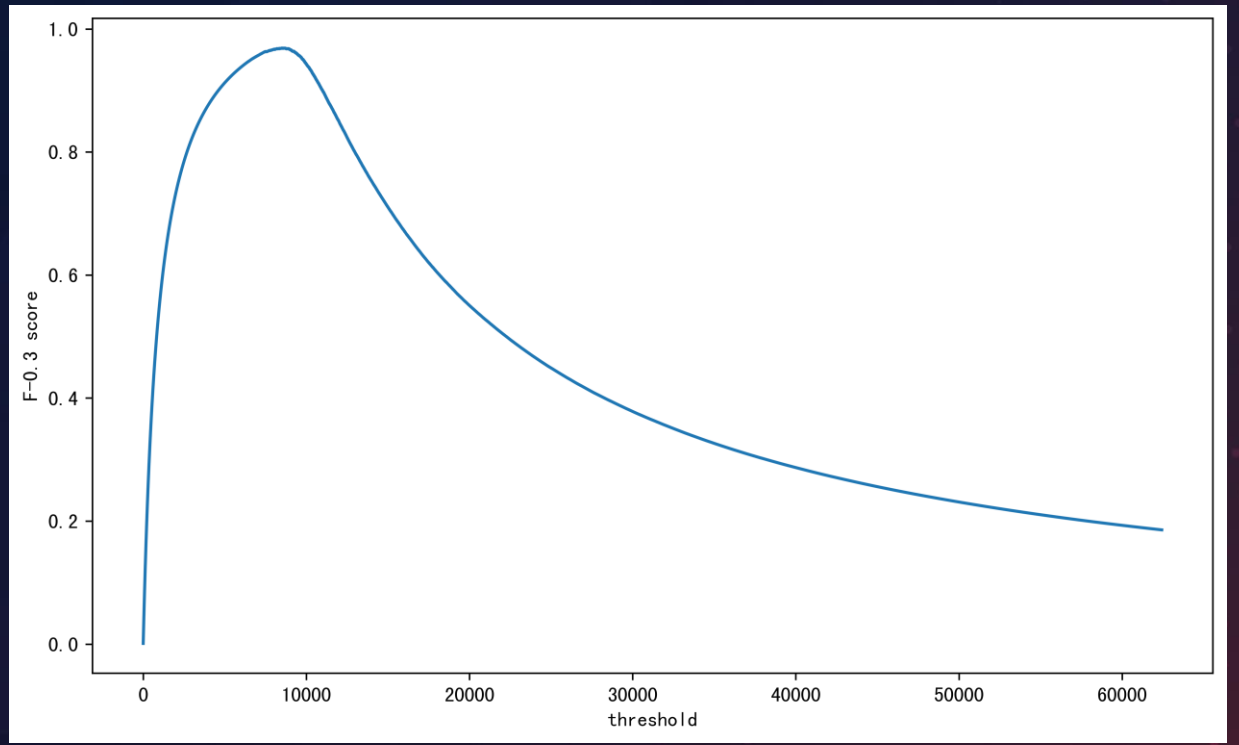
评估指标： $F-beta=(1+beta^2)\frac{Precision*Recall}{beta^2*Precision+Reall}$

模型需要同时兼顾精准率以及召回率两个指标

考虑如何选取最佳阈值（黑样本数量）使得当前模型可发挥最大价值：

- 枚举不同阈值（黑样本数量）在验证集的 F-0.3
- 选取峰值点的阈值比例用于划分测试集的预测结果

最终选用了 2760 个黑样本作为划分依据



字节跳动
安全AI挑战赛

Empower Security
Enrich life

实验相关

- TF-IDF 选取关键字个数（1 个，0.9585、3 个，0.9599、5 个，0.9589）
- 加入个签 / 标题 / POI 是否为空的特征（0.9575 -> 0.9585）
- 加入个签 / 标题长度信息（0.9599 -> 0.9599，效果无变化）
- 加入视频话题数及投稿平均话题数（0.9599 -> 0.9595）
- 加入对个签 / 标题的词袋表征（0.9599 -> 0.9611）
- 对于用户投稿时间平移小时降低训练集与测试集分布差异（0.9599 -> 0.9582）
- 加入个签字母 / 数字 / 符号数量（0.9539 -> 0.9557）
- 加入用户投稿时间中的 weekday 与 hours（0.9531 -> 0.9520）
- 加入用户投稿时间中的 hours（0.9531 -> 0.9527）

5 总结与反思

字节跳动
安全AI挑战赛

字节跳动 安全中心 × 安全范化 BYTEDANCE SECURITY

总结与反思

1. 成绩永无止境

2. 仔细对待特征中的异常情况，谨慎丢弃

3. 感谢主办方提供这次学习的机会，受益匪浅

参考文献

“抖音安全中心协助公安机关破获多起网络色情案件”

https://baijiahao.baidu.com/s?id=1672996855270523715&wfr=spider&for=pc

“一入夜，抖音同城就成了色情入口“

https://www.163.com/dy/article/FMJTD3HT0519E3QB.html

“抖音发布公告：全力打击网络色情、黑色行为 封禁账号超5万“

http://cj.cri.cn/n/20200702/fea5bfed-4673-7fe9-a0cd-c331abbb8566.html

《网络黑产协同治理研究报告》

https://www.docin.com/p-2538025421.html

感谢观看

THANKS FOR WATCHING

字节跳动
安全AI挑战赛

字节跳动 安全中心 × 安全范化 BYTEDANCE SECURITY