2021 字节跳动安全 AI 挑战赛　ECNU_ICA　技术明细

im0qianqian,　Bin,　RR

1. 题目数据

1.1 用户基础信息

字段	说明
id	用户 id
gender_str	性别
signature	个签，按字符映射为 int 值
create_time	创建时间
follow_num_all	关注人数
fans_num_all	粉丝数
publish_cnt_all	投稿数
server_comment_cnt_all	评论数

1.2 用户投稿信息

字段	说明
id	用户 id
item_title	投稿视频标题，按字符映射为 int 值
poi_name	视频 POI，按字符映射为 int 值
item_province_cn	视频省份

item_create_time	视频创建时间

1.3 用户行为信息

字段	说明
id	用户 id
video_play	播放次数
video_play_finish	播放完成次数
play_time	播放时长
click_video_play	点击视频播放
feed_request	feed 请求数
homepage_hot_slide_up	feed 页上滑数
homepage_hot_slide_down	feed 页下滑数
like	点赞数
dislike	点不喜欢数
post_comment	评论数
search	搜索数
share_video	分享数

2. 赛题理解及数据分析

2.1 赛题理解

2.1.1 赛题概况

赛题以网络黑产识别为背景，预测色情导流用户为任务，通过对实际脱敏数据进行预测性分析和评估，并根据实际风控业务的使用场景设立评估标准。

2.1.2 数据概况

用户基本信息与用户行为信息基本为一对一关系，个别用户行为信息缺失；用户基本信息与用户投稿信息为一对多的关系，后续要考虑如何进行转化合并。

2.1.3 判别指标

根据实际风控业务场景，色情导流用户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，故采用 F-beta 对模型进行评估，该计算公式对于预测错误的正常用户会有更大惩罚，以便更加偏向真实使用场景。

2.2 基础数据分析

2.2.1 数据唯一性标准

由于训练集与测试集中包含相同 id，且经主办方确认训练集与测试集中的相同 id 并非同一用户。因此在假设测试集样本标签为未知（-1）时，数据 id 与 label 两列可唯一确定一条样本。

2.2.2 多个数据文件间的关联

题目在训练集与测试集中分别提供了用户基础信息表、用户投稿信息表以及用户行为信息表。此外，训练集额外提供用户标签表以标识训练集中的样本是否属于色情导流用户，若是则 label=1，否则 label=0；假设 label = -1 为未知标签样本，即测试集。

在用户基础信息表中，一行记录代表 id 列所指的用户的基础信息，且 id 唯一。
在用户投稿信息表中，一行记录代表 id 列所指的用户的一条投稿记录，id 不唯一。

在用户行为信息表中，一行记录代表 id 列所指的用户的统计行为信息，且 id 唯一。

其中，用户基础信息表与用户行为信息表根据 id 属于一对一的关系，因此可做直接拼接。
用户基础信息表与用户投稿信息表根据 id 属于一对多的关系，因此可在用户投稿信息中根据 id 聚合并生成有益的统计信息，随后与基础信息表拼接。

3. 特征分析与构建

3.1 特征分析

考虑特征主要分为离散特征，连续特征与文本特征：

- 离散特征（例如：性别、省份等）中类别与类别间一般认为它们之间相互独立，无有序性。
- 连续特征（例如：年龄、粉丝数）中值与值之间存在着明显的有序性，即大小关系。
- 文本特征（例如：个签，视频标题）中含有短网址，联系方式，或诱导性词句，每条数据的该特征之间相互独立。

3.1.1 用户基础特征

在用户基础特征中，主要包括：

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

3.1.2 时间特征

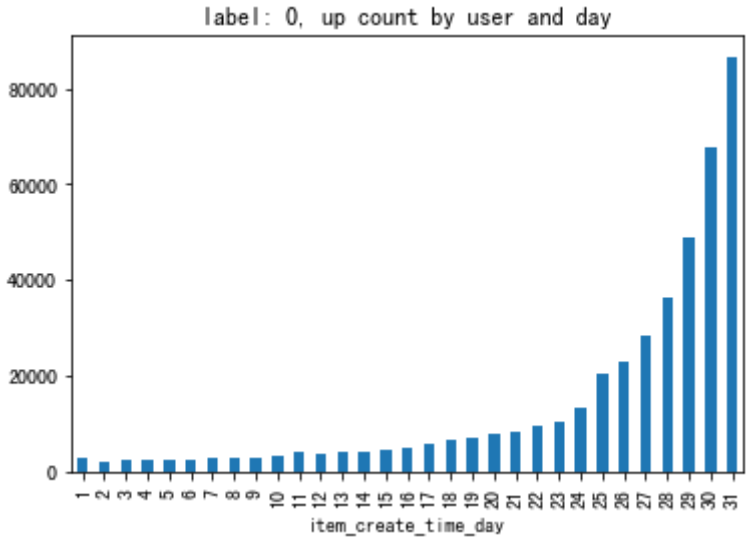
本赛题中时间特征主要有两类：

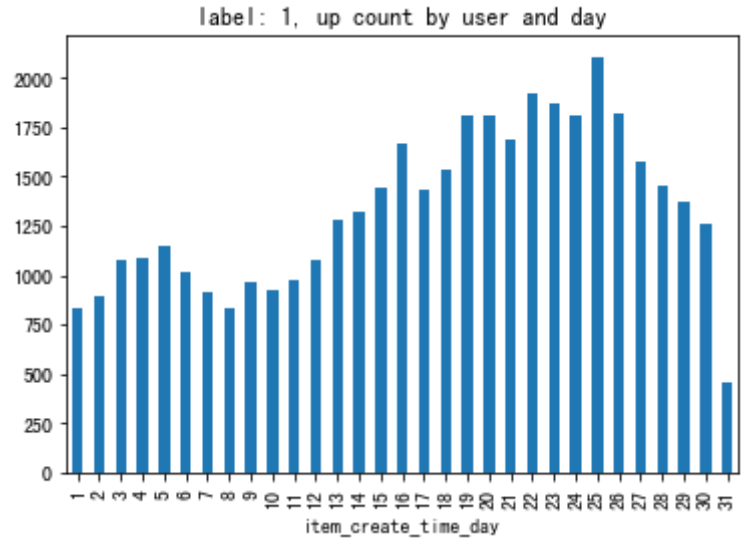
- 用户账户创建时间 create_time
- 视频投稿时间 item_create_time

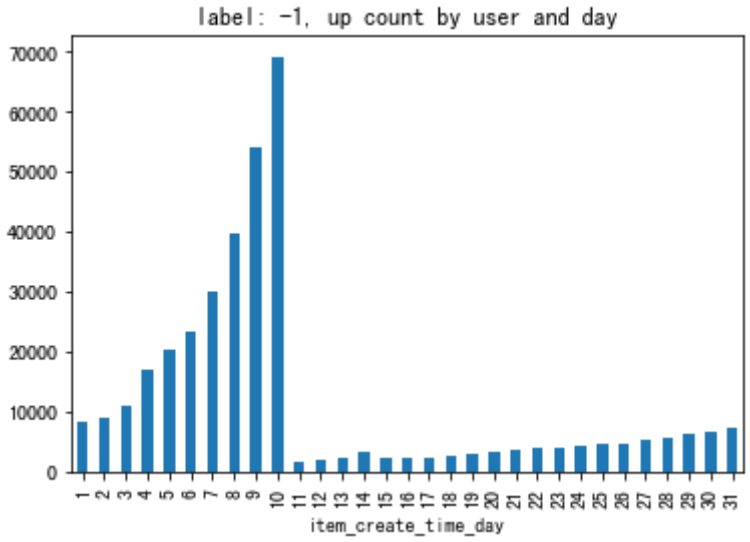
原有的字段为时间戳信息，但可从中提取出年份、月份、天、小时等特征，提取细粒度的时间特征有利于挖掘类似于“色导用户往往在夜间或周末时间段活跃的信息”（举例使用，不保证真实性）。

对训练集中黑白样本以及测试集样本分析可得如下几张图：

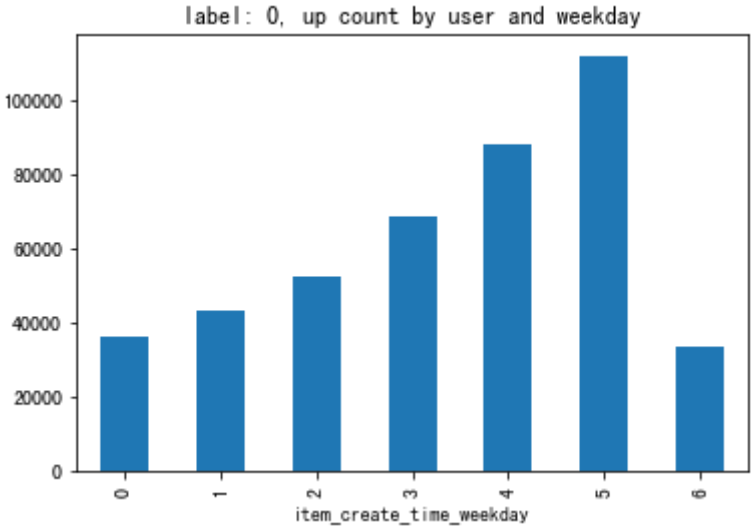
1. 用户投稿在月内 31 天的数量，可以发现黑白样本间的差距非常明显，该特征非常不稳定。

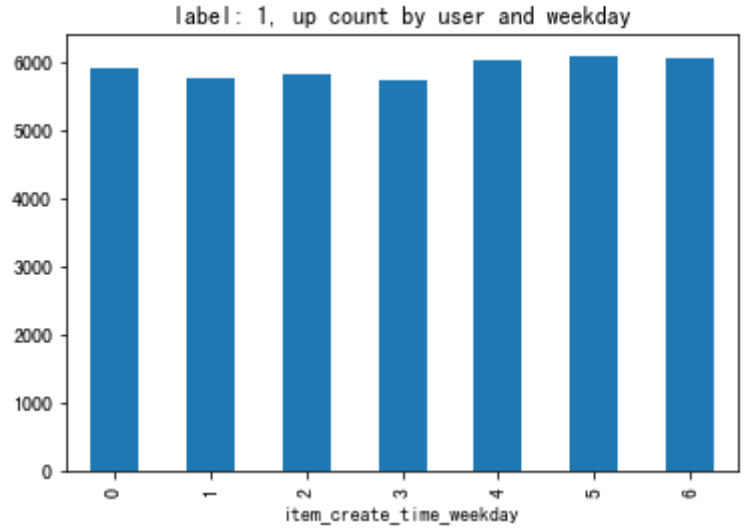


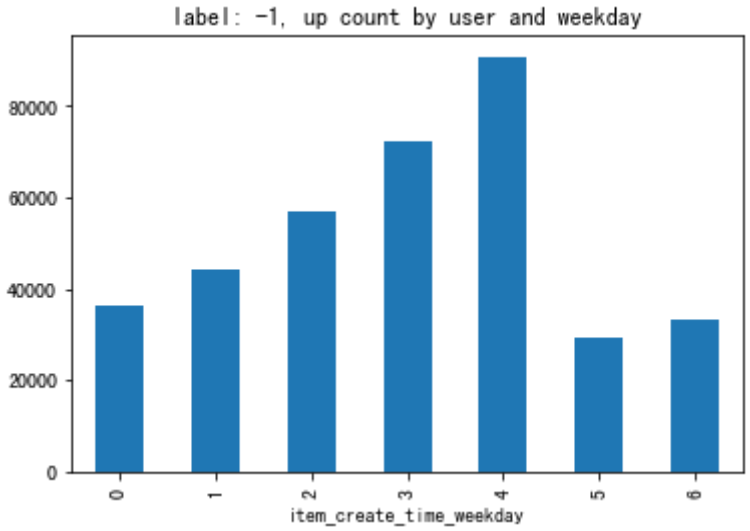




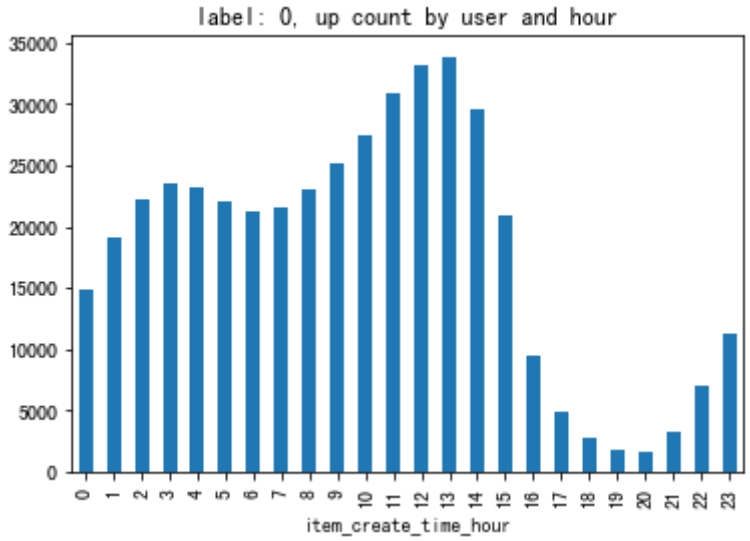
2. 用户投稿在周内七天的数量，可以发现黑样本每天的样本量较为均衡，而白样本周六偏少。

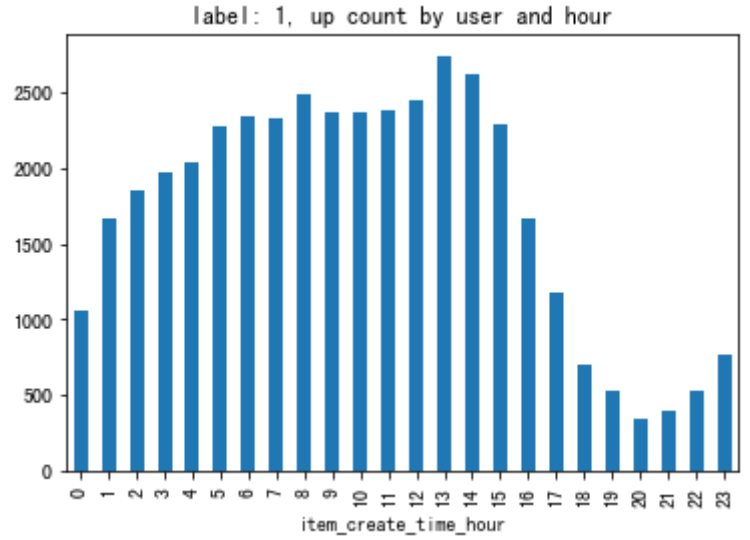


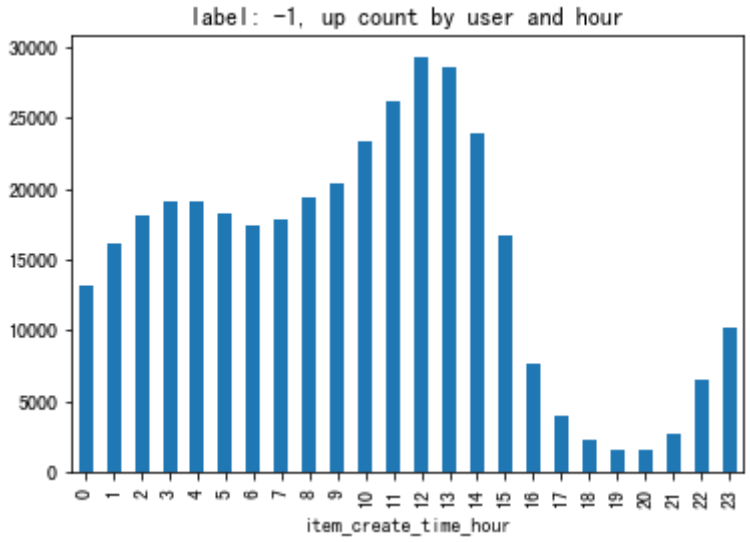




3. 用户投稿在小时内 24 小时的数量，可以发现黑白样本以及测试集数量分布基本一致。







3.1.3 用户行为特征

在用户行为特征中，主要包括：

- video_play，播放次数，连续型特征
- video_play_finish，播放完成次数，连续型特征
- play_time，播放时长，连续型特征
- click_video_play，点击视频播放次数，连续型特征
- feed_request，feed 请求数，但测试集与训练集都为 0，该特征无用，删除
- homepage_hot_slide_up，feed 页上滑数，连续型特征
- homepage_hot_slide_down，feed 页下滑数，连续型特征
- like，点赞数，连续型特征
- dislike，点不喜欢数，连续型特征
- post_comment，评论数，连续型特征
- search，搜索数，连续型特征
- share_video，分享数，连续型特征

3.1.4 字符频率特征

考虑到色情导流用户为了躲避检测，会用一些并不常见即使用频率很低的字符，因此对黑白样本中出现的字符分别做了频率统计，统计仅在 label 为 1 或者为 0 的样本中，个签和视频标题中出现的 index 为独立出现的 index 频率，即在 label 为 1 时出现的 index 而该 index 未曾在 label 为 0 时出现过。

3.2 领域特征挖掘与构建

3.2.1 调研研究

帐号昵称和用户 id 样例

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

个性签名中的联系方式

参考链接：https://www.163.com/dy/article/FMJTD3HT0519E3QB.html

考虑防止被已有模型过滤以及现有样例，色导用户的联系方式开头会通过象形/音译的方式引出联系方式，例如带有 “+v”/“+微”等来比喻微信添加方式，或带有“+q”/“+扣”等来比喻 qq 联系方式，但是也有很多关注度较高的账户也会在个性签名当中留下自己的联系方式来进行商业的导流，所以这并不是具有很强区分度的特征，其中的行为差别就是，正常用户并不会刻意地去使用一些语义上的技巧来隐藏联系方式。我们以此为逻辑点，从调研中发现，常见的色导用户有如下特征：

- 用 emoji 代替文字内容：带有 ♥ v ♥，+等符号来代表微信，加联系方式含义
- 用同音字代替文字内容：加微 =+薇
- 带有“：”联系方式的冒号
- 看头像找我=头像有二维码，扫二维码完成引流

视频 poi 和描述高度相符，通过同城引流

附近的同城用户，同样也是色情导流关注的一个领域。通常黑产团伙会设计好专门的话术，并且加上明显 poi 地点，因为距离的接近，会使得导流的成功大大提升。

- 同城 = 个人名字里带城市名（在 item_province_cn 中的城市)
- poi 方位属于城市/省份之下

视频发布时间

通过之前的调研发现，在抖音的“同城”功能之中，在深夜时段，为色情导流用户的活跃期。他们通过利用人性上的弱点，选择在深夜的时间段进行集中宣发，从而达到导流的目的。调研结果指向了视频发布时间，以此为出发点，考虑投稿时间为深夜时段作为色情导流用户的可能性。

3.2.2 特征构建

个签及投稿标题序列特征

由上述调研研究和赛题描述可发现，个性签名及视频投稿标题中存在色导信息的可能性极大。例如色导用户常用一些 emoji 符号来代替一些文字，而普通用户则不会这样做，这使得色导用户个签信息中存在着标志性的字符。

因此，本工作考虑针对此文本序列信息进行建模，具体的操作有：

1. 将用户个签及所有投稿视频的标题拼接以形成该用户的描述性文本
2. 基于 TF-IDF 算法从该描述性文本中抽取前 3 个代表性的关键字（测试过关键字数量对于模型效果的影响，3 个关键字效果优于 1 个，也优于 5 个）
3. 基于词袋模型为描述性文本进行编码，且取出现频次最高的 2500 个字符所对应的维度

黑白样本对比情况下，个签中不重复出现的字符

由上述调研研究可得知，对于色情导流用户在文本特征导流时，会通过特殊字符/谐音字/象形字来规避当前的审查机制，从而使用一些平日不经常用的字词或者符号，一定程度上迷惑当前模型。以此为出发点，考虑在不同的标签下，对应群体的个签及视频标题中是否有相对的生僻字/相对使用较少的符号。并统计仅在 label 为 1 或者为 0 的样本中，个签和视频标题中出现的 index 为独立出现的 index 频率，即在 label 为 1 时出现的 index 而该 index 未曾在 label 为 0 时出现过（例如对于用户个签中出现过的唯一值，在 label 为 0 的样

本中出现过 index:2693，而在 label 为 1 的样本中并没有出现过该 index）

视频投稿发布时间（小时）特征

由上述调研研究可发现，视频发布时间（小时）或许对色导用户的识别有着正向的帮助。但实际测试发现，视频发布时间特征（小时）对本地验证集中效果有提升，而线上效果略微降低，猜测由于测试集与训练集在视频发布时间（小时）中的分布存在着些许差异。

明文链接特征

由上述调研总结与猜测，色导用户个签及所投稿视频标题中或许存在链接以导流用户到其他视频 / 平台。因此考虑根据规则鉴别链接信息，该规则为匹配满足 https:// 或 http:// 的子串并发现脱敏表中与 https:// 所对应的字符。进而可用于标识文本中的链接数量。
基于该规则，可检测到脱敏 id 与字符间有如下对应关系：

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

个签 / 标题 / POI 是否为空的特征

根据训练集分析发现大量正常用户的个签列为空，而个签作为序列信息难以表述其空/非空，因此加入该特征。在特征构建中实际同时引入 个签、视频标题、POI 是否为空的特征，最终测试证明该信息是有用的（0.9575 -> 0.9585）。

个签中的字母 / 数字 / 符号数量的特征

在确定了 http 与 https 的特征之后，我们将个签中可能存在的 url 链接提取了出来，进行了

进一步地分析。因为字符在 url 中出现的频率和顺序并不会改变，通过对一级域名以及最后一级域名的特征进行对比，以及很多轮的信息迭代，最终确定了 url 中存在的字符、数字以及部分的符号，最后按照组合特征加到了模型当中。

3.3 特征交叉与聚合

3.3.1 聚合特征

根据 id 聚合投稿信息表

由于用户基础信息表及用户标签表与用户投稿信息表是一对多的关系，即同一个用户在投稿信息表中关联多条数据。因此考虑根据用户 id 聚合投稿信息表，并统计出一些关键特征用于增强描述单一用户的画像表示。
所谓聚合操作（groupby）即根据指定列（例如 id）进行聚合，对于离散列 id 中的每一个类别，求出其他列的一些聚合特征，例如某一列的平均值（mean）、最大值（max）、数量（count）等。

在本实验中，投稿信息表主要包含以下字段：

- 用户 id，离散类，非特征
- item_title，投稿视频标题，序列特征
- poi_name，视频 POI，序列特征（也可当作离散特征）
- item_province_cn，视频省份，离散特征
- item_create_time，视频创建时间，连续特征

在经由前文处理，新增特征如下：

- is_item_title_contain_url，视频标题中所含链接（http + https）数量，连续特征
- is_item_title_null，视频标题是否为空，离散特征
- is_poi_name_null，视频 POI 信息是否为空，离散特征

在聚合操作中，选取用户 id 作为聚合索引，对于视频标题（item_title）、POI 信息（poi_name）以及视频省份（item_province_cn）聚合求得 nunique（不同值的数量）以及 count（该 id 下所关联的数据个数）。对于视频创建时间（item_create_time）、视频标题链接数量（is_item_title_contain_url）、视频标题是否为空（is_item_title_null）以及视频 POI 信息是否为空（is_poi_name_null）聚合求得其中位数、平均值、最大值、最小值、求和、方差以及标准差信息。

需要注意的是，由于用户投稿时间在训练集与测试集中分布差异巨大，因此考虑对于视频创建时间仅保留一些与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。

根据基础离散特征做聚合

考虑当前缺乏以类别特征为依据并做聚合的相关信息，该聚合操作有利于基于类别列筛选而挖掘出该类别相关的独有信息。例如性别为男的用户的平均账户创建时间等。
类似的，该聚合也可在多个类别列间进行组合而产生。例如性别为男且个性签名为空的用户的平均粉丝数是多少。

设定特征离散列分别为：

- gender_str，性别
- signature，个性签名
- tfidf_keyword_0，前文中基于个签及标题序列特征所抽取的 tfidf 第一关键字
- is_signature_null，个性签名是否为空

设定特征连续列分别为：

- create_time，账户创建时间
- follow_num_all，关注人数
- fans_num_all，粉丝数
- publish_cnt_all，投稿数
- server_comment_cnt_all，评论数
- video_play，播放次数
- video_play_finish，播放完成次数
- play_time，播放时长
- click_video_play，点击视频播放次数
- homepage_hot_slide_up，feed 页上滑数
- homepage_hot_slide_down，feed 页下滑数
- like，点赞数
- dislike，点不喜欢数
- post_comment，评论数
- search，搜索数
- share_video，分享数

- create_time_year，账户创建时间（年）
- create_time_month，账户创建时间（月）
- create_time_day，账户创建时间（日）
- create_time_weekday，账户创建时间（周几）
- create_time_hour，账户创建时间（小时）
- is_signature_contain_url，个签所包含的链接数量
- sparse:user_up_info:group_val_(id)_item_title_nunique，用户投稿中标题不同的视频数量
- sparse:user_up_info:group_val_(id)_poi_name_nunique，用户投稿中 POI 不同的视频数量
- sparse:user_up_info:group_val_(id)_item_province_cn_nunique，用户投稿中视频省份不同的数量
- sparse:user_up_info:group_val_(id)_count，用户投稿表中的记录数
- dense:user_up_info:group_val_(id)_item_create_time_(max-min)/count，用户平均投稿时间
- dense:user_up_info:group_val_(id)_item_create_time_(max-min)，用户投稿跨度
- dense:user_up_info:group_val_(id)_is_item_title_contain_url_mean，用户投稿标题包含链接数的平均值
- dense:user_up_info:group_val_(id)_is_item_title_contain_url_sum，用户投稿标题包含链接数的和

3.3.2 交叉特征

考虑特征之间的交叉容易发现一些更为有用的新特征，例如播放时长 / 播放次数可以得出平均播放时长；粉丝数 / 投稿数可以得出平均每个投稿所获粉丝量。

因此，本次实验将基于前文“根据基础离散特征做聚合”中的连续列之间做特征交叉，交叉的方式主要包含：

- A * B
- A + B
- A - B
- A / (B + 0.00000001)

实验也验证了该方法的有效性。

4. 模型构建

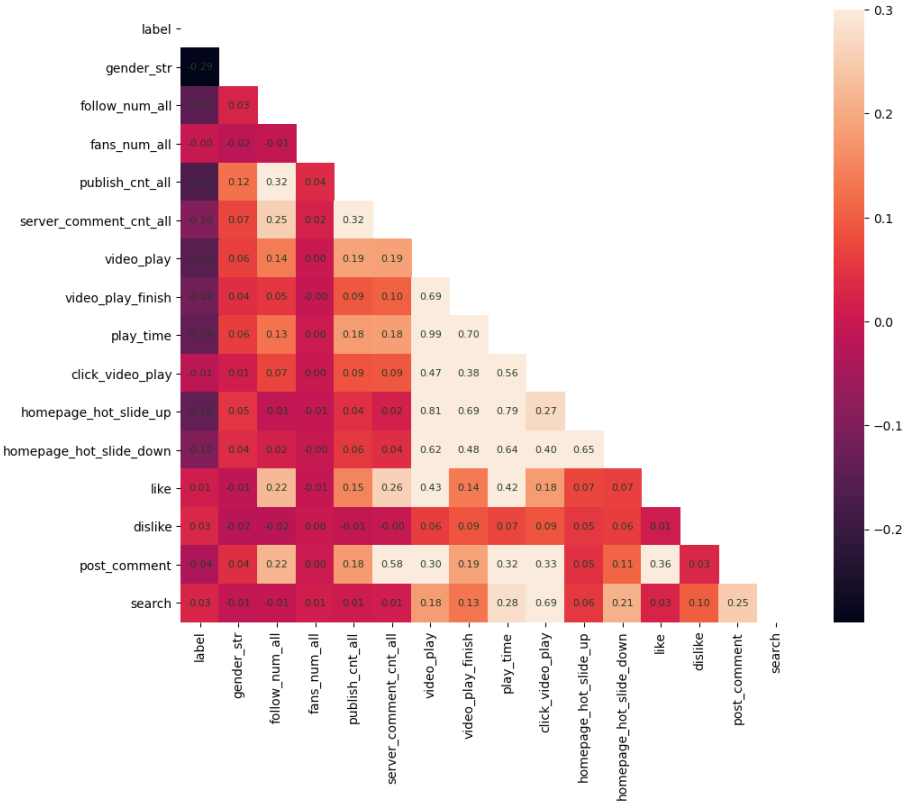
4.1 类别特征 id 化

该操作主要是由于下游模型可能存在不接受以非整数 / 浮点数所描述的特征，因此程序在现有构建的特征基础上做类别特征 id 化。

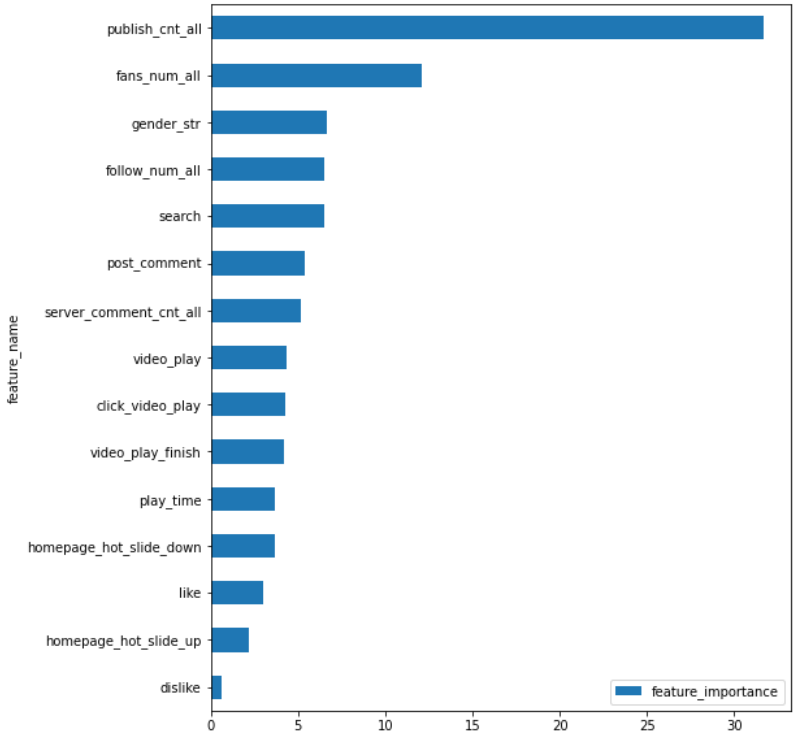
其大致思想要为一列特征中每一个类别分配一个唯一的数字进行代替。其他也有类似于 one-hot 的方案。

4.2 决策树模型设立和调参

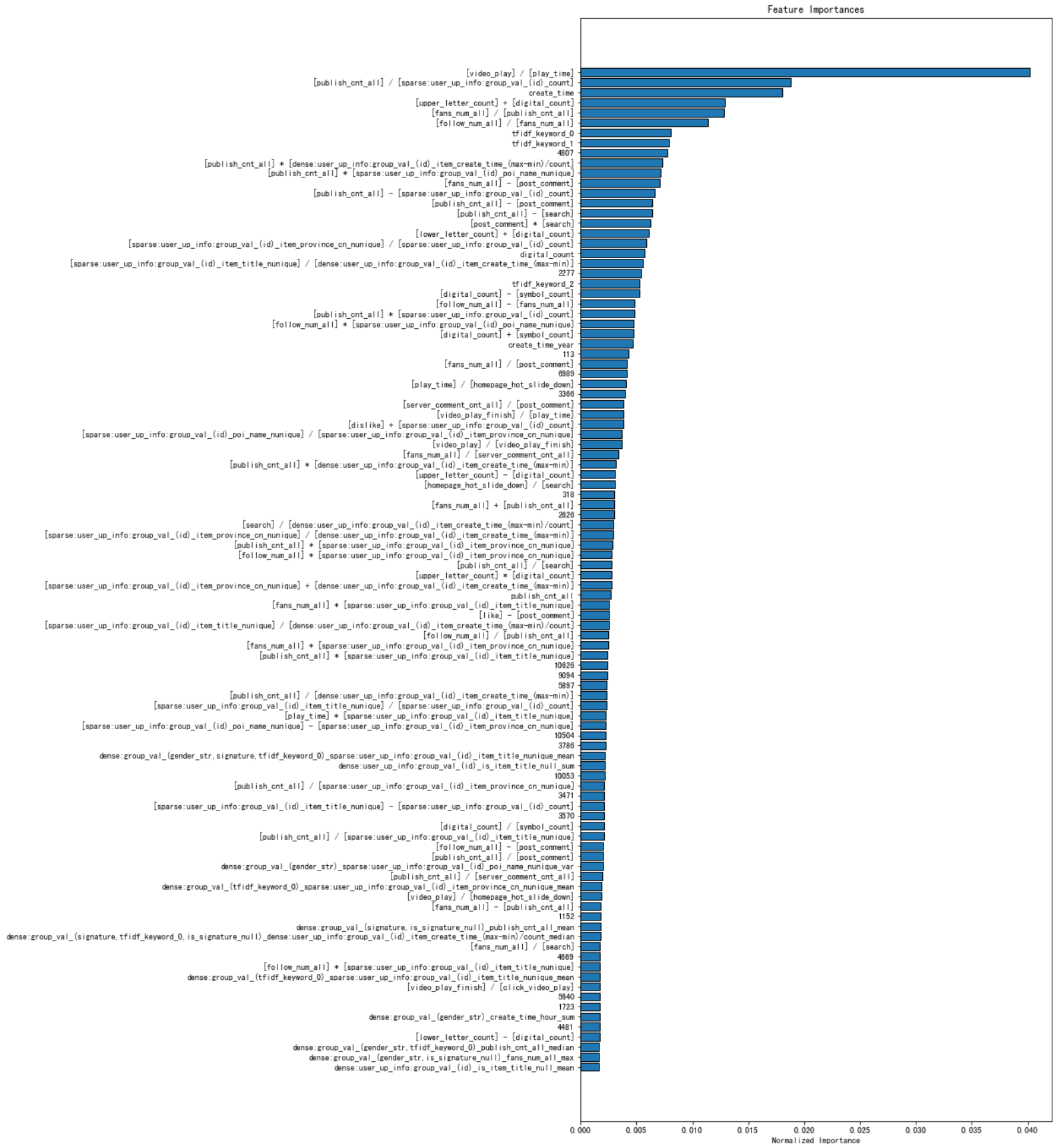
- 模型尝试：lightgbm，xgboost，catboost，最终选择 catboost 作为 backbone model
- 模型调优：在训练集中做五折交叉验证，其中每一折训练保留验证集中具有最大 fbeta-0.3 的模型迭代
- 基础特征相关性：根据皮尔逊相关系数计算已有基础特征及 label 之间的相关性，并绘制热力图如下。



- 基础特征重要性：将基础特征送入 catboost 训练所得基础特征重要性如下图所示。



- 完整特征重要性：根据 catboost 模型训练结果查看完整特征重要性，保留重要的部分特征，用于特征筛选。下图即为本实验所得特征重要性 Top 100 排行。


Feature Importances
[video_play] / [play_time]
[publish_cnt_all] / [sparse:user_up_info:group_val_(id)_count]
create_time
[upper_letter_count] + [digital_count]
[fans_num_all] / [publish_cnt_all]
[follow_num_all] / [fans_num_all]
tfidf_keyword_0
tfidf_keyword_1
4807
[publish_cnt_all] * [dense:user_up_info:group_val_(id)_item_create_time_(max-min)/count]
[publish_cnt_all] * [sparse:user_up_info:group_val_(id)_poi_name_nunique]
[fans_num_all] - [post_comment]
[publish_cnt_all] - [sparse:user_up_info:group_val_(id)_count]
[publish_cnt_all] - [post_comment]
[publish_cnt_all] - [search]
[post_comment] * [search]
[lower_letter_count] + [digital_count]
[sparse:user_up_info:group_val_(id)_item_province_cn_nunique] / [sparse:user_up_info:group_val_(id)_count]
digital_count
[sparse:user_up_info:group_val_(id)_item_title_nunique] / [dense:user_up_info:group_val_(id)_item_create_time_(max-min)]
2277
tfidf_keyword_2
[digital_count] - [symbol_count]
[follow_num_all] - [fans_num_all]
[publish_cnt_all] * [sparse:user_up_info:group_val_(id)_count]
[follow_num_all] * [sparse:user_up_info:group_val_(id)_poi_name_nunique]
[digital_count] + [symbol_count]
create_time_year
113
[fans_num_all] / [post_comment]
6989
[play_time] / [homepage_hot_slide_down]
3366
[server_comment_cnt_all] / [post_comment]
[video_play_finish] / [play_time]
[dislike] + [sparse:user_up_info:group_val_(id)_count]
[sparse:user_up_info:group_val_(id)_poi_name_nunique] / [sparse:user_up_info:group_val_(id)_item_province_cn_nunique]
[video_play] / [video_play_finish]
[fans_num_all] / [server_comment_cnt_all]
[publish_cnt_all] * [dense:user_up_info:group_val_(id)_item_create_time_(max-min)]
[upper_letter_count] - [digital_count]
[homepage_hot_slide_down] / [search]
318
[fans_num_all] + [publish_cnt_all]
2626
[search] / [dense:user_up_info:group_val_(id)_item_create_time_(max-min)/count]
[sparse:user_up_info:group_val_(id)_item_province_cn_nunique] / [dense:user_up_info:group_val_(id)_item_create_time_(max-min)]
[publish_cnt_all] * [sparse:user_up_info:group_val_(id)_item_province_cn_nunique]
[follow_num_all] * [sparse:user_up_info:group_val_(id)_item_province_cn_nunique]
[publish_cnt_all] / [search]
[upper_letter_count] * [digital_count]
[sparse:user_up_info:group_val_(id)_item_province_cn_nunique] + [dense:user_up_info:group_val_(id)_item_create_time_(max-min)]
publish_cnt_all
[fans_num_all] * [sparse:user_up_info:group_val_(id)_item_title_nunique]
[like] - [post_comment]
[sparse:user_up_info:group_val_(id)_item_title_nunique] / [dense:user_up_info:group_val_(id)_item_create_time_(max-min)/count]
[follow_num_all] / [publish_cnt_all]
[fans_num_all] * [sparse:user_up_info:group_val_(id)_item_province_cn_nunique]
[publish_cnt_all] * [sparse:user_up_info:group_val_(id)_item_title_nunique]
10626
9094
5897
[publish_cnt_all] / [dense:user_up_info:group_val_(id)_item_create_time_(max-min)]
[sparse:user_up_info:group_val_(id)_item_title_nunique] / [sparse:user_up_info:group_val_(id)_count]
[play_time] * [sparse:user_up_info:group_val_(id)_item_title_nunique]
[sparse:user_up_info:group_val_(id)_poi_name_nunique] - [sparse:user_up_info:group_val_(id)_item_province_cn_nunique]
10504
3786
dense:group_val_(gender_str,signature,tfidf_keyword_0)_sparse:user_up_info:group_val_(id)_item_title_nunique_mean
dense:user_up_info:group_val_(id)_is_item_title_null_sum
10053
[publish_cnt_all] / [sparse:user_up_info:group_val_(id)_item_province_cn_nunique]
3471
[sparse:user_up_info:group_val_(id)_item_title_nunique] - [sparse:user_up_info:group_val_(id)_count]
3570
[digital_count] / [symbol_count]
[publish_cnt_all] / [sparse:user_up_info:group_val_(id)_item_title_nunique]
[follow_num_all] - [post_comment]
[publish_cnt_all] / [post_comment]
dense:group_val_(gender_str)_sparse:user_up_info:group_val_(id)_poi_name_nunique_var
[publish_cnt_all] / [server_comment_cnt_all]
dense:group_val_(tfidf_keyword_0)_sparse:user_up_info:group_val_(id)_item_province_cn_nunique_mean
[video_play] / [homepage_hot_slide_down]
[fans_num_all] - [publish_cnt_all]
1152
dense:group_val_(signature,is_signature_null)_publish_cnt_all_mean
dense:group_val_(signature,tfidf_keyword_0,is_signature_null)_dense:user_up_info:group_val_(id)_item_create_time_(max-min)/count_median
[fans_num_all] / [search]
4669
[follow_num_all] * [sparse:user_up_info:group_val_(id)_item_title_nunique]
dense:group_val_(tfidf_keyword_0)_sparse:user_up_info:group_val_(id)_item_title_nunique_mean
[video_play_finish] / [click_video_play]
5640
1723
dense:group_val_(gender_str)_create_time_hour_sum
4481
[lower_letter_count] - [digital_count]
dense:group_val_(gender_str,tfidf_keyword_0)_publish_cnt_all_median
dense:group_val_(gender_str,is_signature_null)_fans_num_all_max
dense:user_up_info:group_val_(id)_is_item_title_null_mean
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
Normalized Importance

4.3 PR 曲线及阈值选择

本次竞赛采用 F-0.3 作为评估标准，如下式：

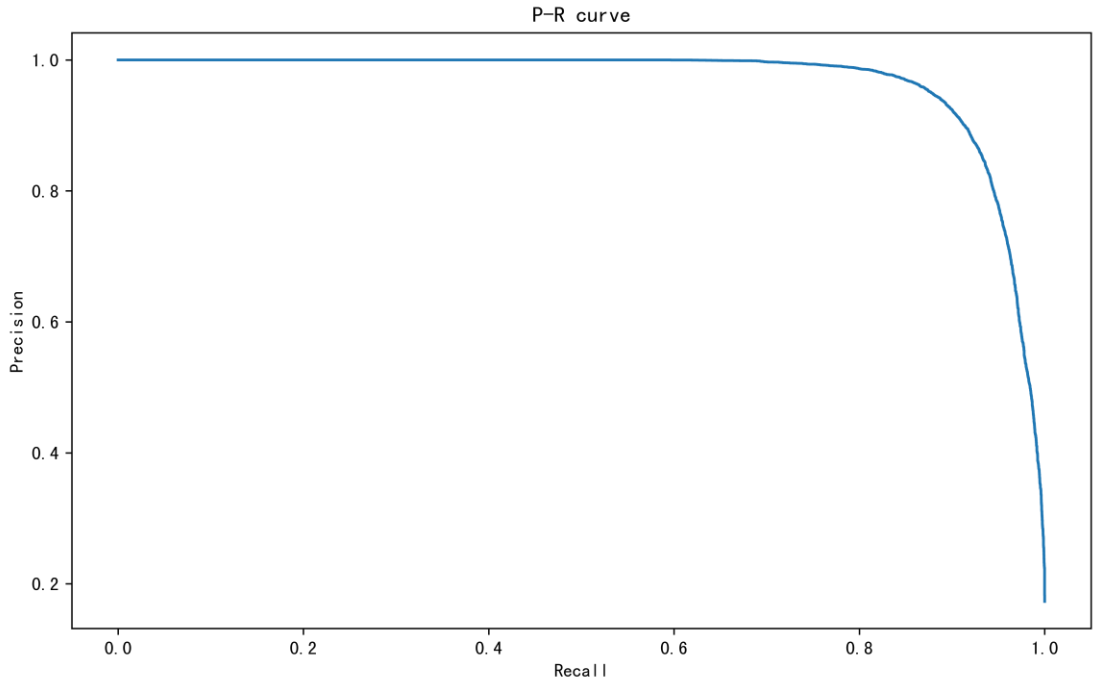
$$F-beta=(1+beta^2)\frac{Precision*Recall}{beta^2*Precision+Reall}$$

可见，模型需要同时兼顾精准率以及召回率两个指标。

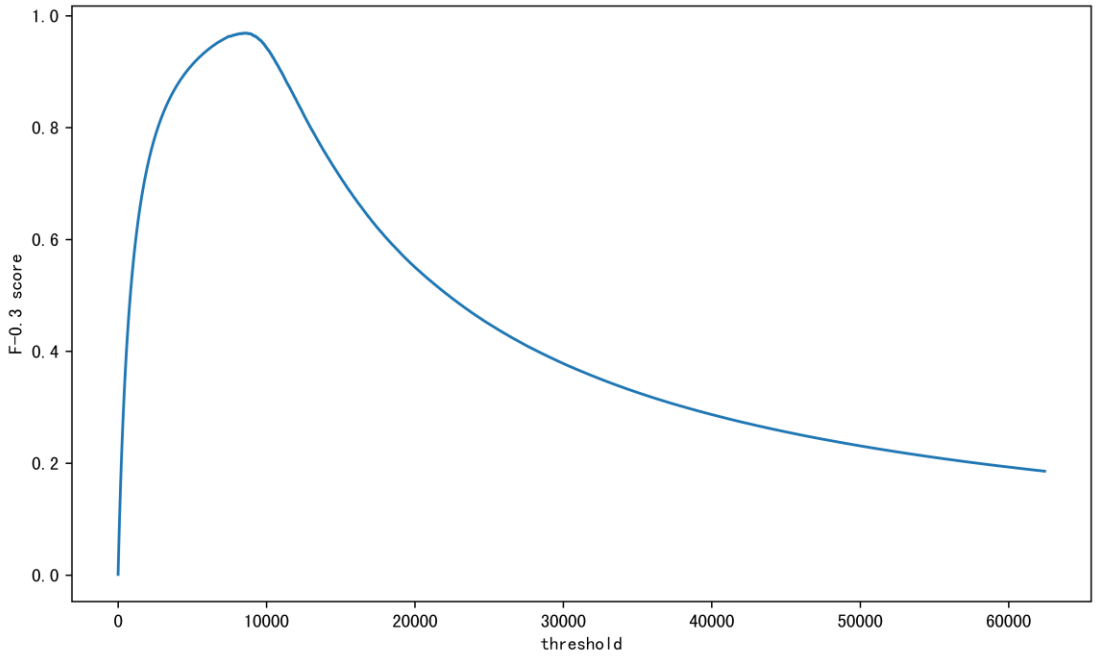
在此类问题中，模型的作用即为每一个样本进行打分，打分结果接近 1 的更有可能为色导用户。也有评估指标 AUC 用于衡量模型对黑样本打分高于白样本的能力。在所有样本中，应该决策有多少个样本属于黑样本且可以达到最高的评估标准也是一个需要探讨的点。

因此，本实验尝试绘制 P-R 曲线以及不同阈值下在验证集中的 f-0.3 指标，发现了如下现象：

P-R 曲线



不同阈值（黑样本数量）在验证集的 F-0.3 score



实验发现随着阈值的上升，总有一个点的阈值可以为验证集 F-0.3 score 指标达到峰值，该阈值点均衡了精准率以及召回率，且可以作为预测测试集阈值的重要参考。而在该阈值左侧梯度较小，右侧梯度较大，这也可以为寻找测试集最佳阈值提供有力的指导。
最终我们选用了 2760 作为预测结果中黑样本的数量，于复赛取得了 0.9611 的成绩。

5. 所做过的尝试以及思考

- 将每个用户最早和最近上传视频的日期作为特征时，验证集效果非常好，但上线后成绩很差，从而发现训练集与测试集时间分布差距过大。
- 构建特征相关系数矩阵热力图，计算特征重要程度分数。
- 通过调查相关资料进行导流用户分析和判断，建立可能相关的特征，包括投稿视频发布时间，个签和标题中网址数量。
- 投稿时间对于 index 的影响，根据调查相关资料得知深夜时段可能存在较多的色情导流用户投稿，构建对应的特征。
- 对于平均投稿时间间隔，投稿时间进行归一化处理，但训练集与测试集时间分布差距过大。
- 通过个签和视频标题推断字母/数字/符号可能的 index。
- 加入个签，标题长度作为特征。
- 考虑到题目描述“并且通过二维码、联系方式、短网址等完成导流”，考虑连续字母+数字的组合，或者较为明显的联系方式特征（+qq：后带连续数字，+vx：后带数字+字

母+符号组合等）。

- 考虑到+qq，+vx 这种特征可能，尝试通过深度学习构建序列信息模型。
- 统计仅在 label 为 1 或者为 0 的样本中，个签和视频标题中出现的重复的 index，判断可能没有帮助的且反复出现次数过多的 index。
- 对于 1 和 0 的 label 查找仅在 1 中出现的 index 以及仅在 0 中出现的 index，统计测试集中每个 id 下的个签和视频标题中，对应仅在 1 或者 0 出现的 index 个数，作为特征。
- 对于出现过的 index 进行 one-hot 编码，将所有个签 padding 到相同长度，并对出现的 index 进行 one-hot 编码。但这会导致每个特征维度过大，且特征过于稀疏，对于模型来说不具有足够的学习能力，并且会大幅增加训练模型所需时间。
- 针对可能出现的个签重复话语次数和长度进行统计，并作为特征。但是未能实现最大长度下统计重复次数的方案，转为指定长度重复序列。通过统计 3/4 个 index 值的和，进行每个个签内部的匹配，并记录重复次数，通过次数/总长度判断重复度。在训练集两种 label 分别统计中，该特征量上具有较为明显的差距，但作为特征实际训练模型时效果下降。