小白 UP 主打怪升级路线——bilibili 视频热度分析

小组成员：李云帆、施俞晨、邓淇升、王玉世

一、背景介绍

Bilibili 是一个基于弹幕视频分享的互联网社区，于 2009 年成立。该网站发展迅速，目前已成为 ACG[1]、二次元领域的头部网站之一。2017 年 7 月，网站月活跃用户量为 5843 万，相较去年同期增长了 63.2%，日新增用户量为 45.7 万。该网站用户以大学生为主，男生比例略多于女生，主要关注音乐、书籍、二次元和校园等内容。[i]该网站主要分为动画、番剧、音乐、舞蹈、游戏等十个板块，每个板块下面又细分了不少子领域，用户可以方便地通过不同分区找到自己心仪的内容。

用户如何表达认可？最基础的是“观看”，因此视频的播放量是一个重要指标。同时，B 站用户也可以通过点赞、评分、推荐、收藏、投硬币、充电等方式[2]表达自己对视频内容的认可。

大家也知道我们复旦今年刚毕业的校友老番茄[3]是 B 站的著名 UP 主，在最近的毕业典礼上主持复旦新闻联播，大放异彩。我们希望以他为榜样，研究如何做成功 UP 主。

在本研究中，我们希望从一个新进入 UP 圈的小白 UP 主[4]出发，考察各个变量与播放量的关系，为其提升播放，提高人气提供一定参考意见，后面的文本聚类分析为小白 UP 主选择切入领域提供了。本文接下来将分为以下几个部分：（二）数据处理及描述性统计，（三）回归（四）聚类模型（五）总结、建议及反思 （六）附录——一些有趣的工作

二、描述性统计

我们的数据库中包含了我们爬取的 101 位知名 UP 主的 28000 多条视频的数据。我们选择了其中播放量、频道、点赞、投币、收藏、分享等变量进行建模。原始变量表如下：

---

[1] ACG = Animation 动画 + Comic 漫画 + Game 游戏

[2] B 站也会根据点赞、收藏等数据为视频计算热度，并据此决定将哪些视频置入首页推荐给所有用户观看，因此点赞、播放等不仅表达了时下用户对视频的认可度，更在一定程度上决定了视频未来能够具有的影响力。

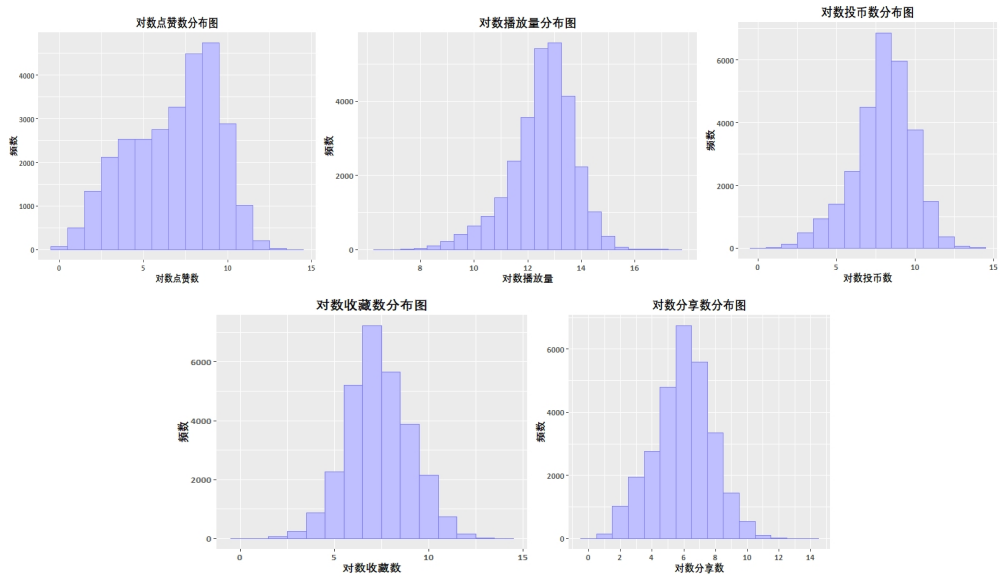
[3] 复旦金融系大四学生，著名 B 站 UP 主，知名网络播客和游戏娱乐解说人，粉丝数：367 万，自制视频总播放量：2.9 亿，成名作：《你可见过如此丧心病狂的口袋妖怪解说》系列。

[4] UP 主是制作视频并将其上传到网站中的人。UP 主生产视频内容，普通用户观赏、消费这些内容。更加获得用户认可的 UP 主则具有更高人气，一方面这为 UP 带来了成就感和满足感，另一方面这些巨大的人气也极易通过打广告等方式变现。因此，如何才能让更多用户认可，如何才能拥有更多粉丝，是许多 UP 十分重视的问题。

变量类型	变量名	变量说明	取值范围
因变量	**播放量**	离散变量	整数：817 – 37128221，单位：次
自变量（视频信息）	视频标题	文本数据	例：A路人做声线测试，最后疯了
	视频时长	离散变量	整数：0 – 17129，单位：秒
	视频号	分类变量（28459水平）	每个视频的视频号独一无二，例：8741476
	频道	分类变量（60水平）	例：单机游戏、美食圈、手机平板等
	标签	文本数据	例：#攻略##黑桐谷歌##PS4#
	点赞	离散变量	整数：0 – 1528233，单位：次
	投币	离散变量	整数：0 – 1944322，单位：个
	收藏	离散变量	整数：0 – 1423719，单位：次
	分享	离散变量	整数：0 – 873886，单位：次

我们的因变量为播放量[5]。我们对播放量进行了对数化处理。除了播放量之外，“三连”[6]，即点赞、投币、收藏，也显示了用户的认可，三个变量连同‘分享’变量，最大值都出现在视频《改革春风吹满地》。

以下是对数化以后的各个变量的对数分布直方图。



[5] 最小播放量为 817，最大为 3700 万。播放量最小的视频是 UP 主起小点是大腿的《L 星人 snake 经理左雾：就算 0.1%的帮助我也会去做》，最大播放量为小可儿的《改革春风吹满地》。

[6] 三连：点赞 + 投币 + 收藏

下图是不同变量的基本统计量信息：

视频人气指数	结果
最小值	1.7
最大值	14.4
均值	7.5
中位数	7.6

对数点赞数	结果
最小值	0
最大值	14.2
均值	6.8
中位数	7.3

对数投币数	结果
最小值	0
最大值	14.5
均值	8.0
中位数	8.1

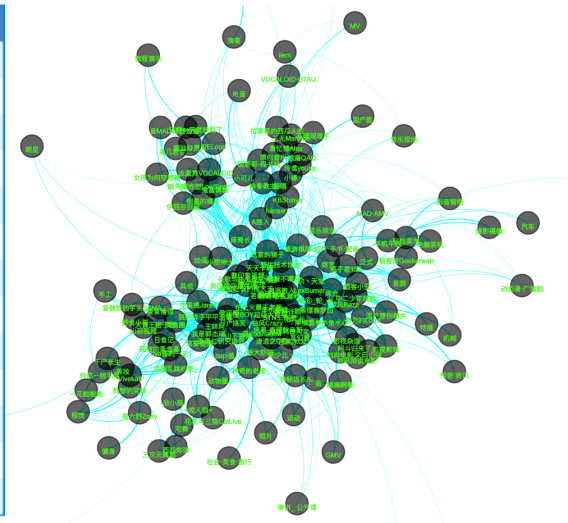
对数收藏数	结果
最小值	0
最大值	14.2
均值	7.4
中位数	7.3

对数分享数	结果
最小值	0
最大值	13.7
均值	5.9
中位数	6.0

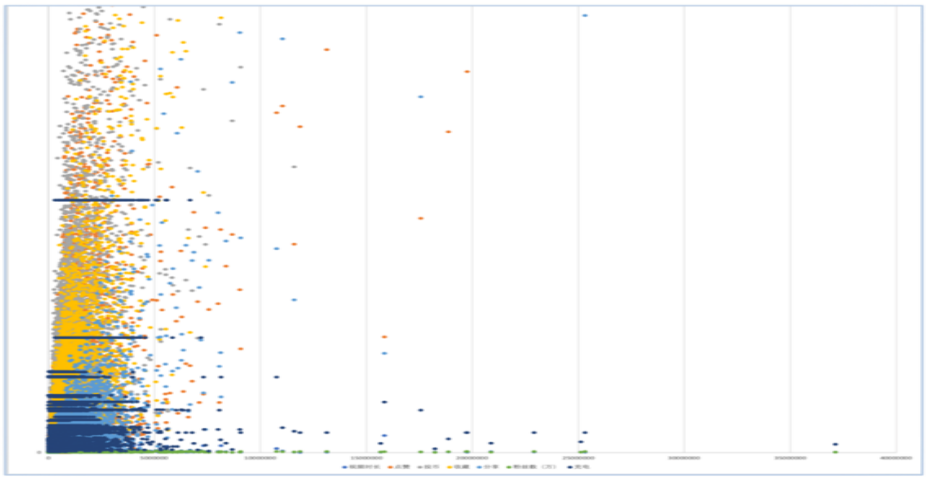
重要变量：频道，共有 60 个视频频道，

超过 500 个视频的频道分为如下左图所示的 13 个类，下右图为 UP 们和频道的社群图

频道	视频数
单机游戏	6877
电子竞技	4410
搞笑	2451
美食圈	1871
影视杂谈	1690
日常	1576
综合	1431
翻唱	1292
网络游戏	1108
鬼畜调教	988
美妆	855
手机游戏	630
手机平板	508



观察下图播放量关于其他变量的散点图可知：



1.各个变量在播放量较低处较为集中，但又有播放量与点赞，投币，分享这几个变量有强正相关。

2. 粉丝数在我们的观察中比较稳定在百位数量级。

3. 视频时长也相对来说波动较小

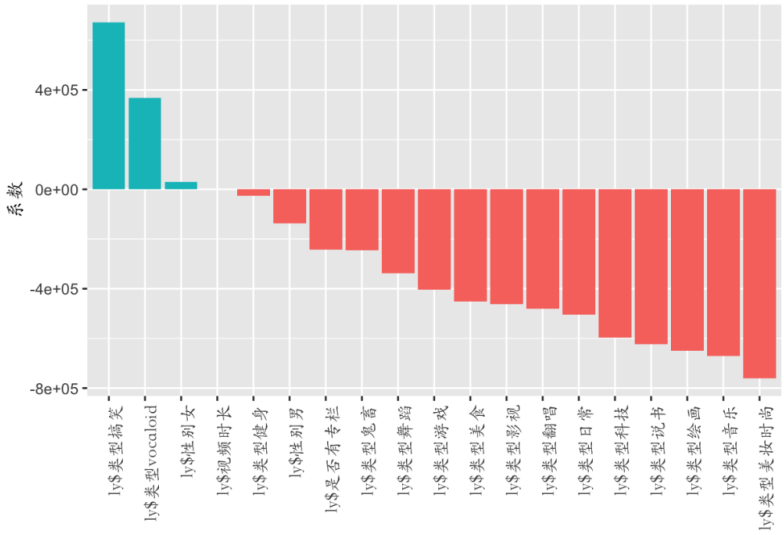
接下来根据上述观察进行回归分析

三、回归模型

在回归模型部分，我们考察了线性回归、LASSO、回归树和 xgboost。

各个回归模型结果如下：

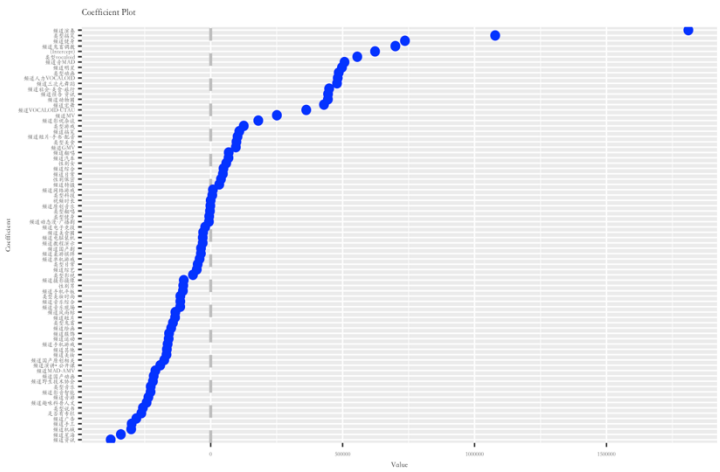
线性回归：



通过线性回归的结果我们可以发现，类型中搞笑和 vocoid 会增加播放量，而健身、鬼畜等会降低播放。性别的系数较小，但相较于保密性别，女生会降低播放，而男生会增加。有专栏会增加播放，视频时长似乎没有明显影响。

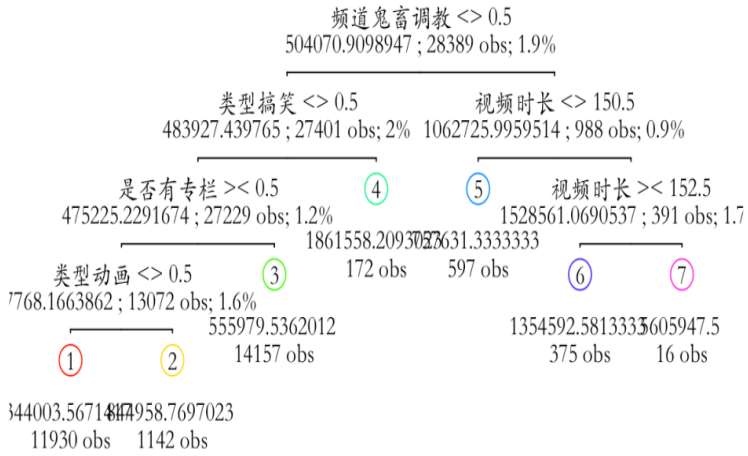
在上一个线性回归的基础上，我们加入了频道这一变量。发现演奏、健身、鬼畜、明星等频道都会增加播放，但是资讯、星海、机械、运动等会降低播放量。

LASSO 回归：



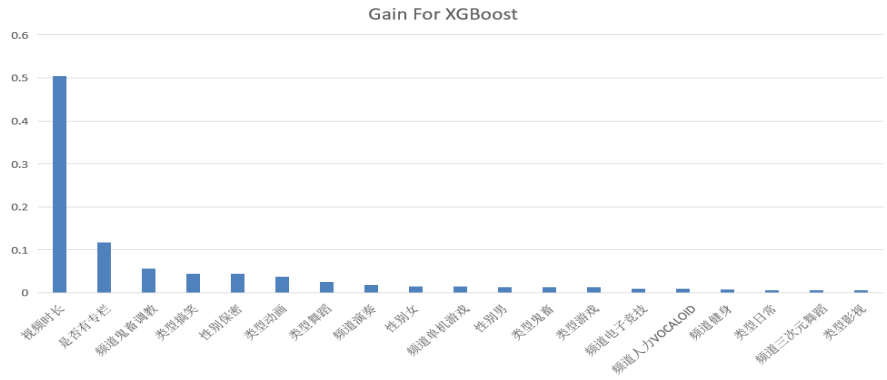
我们发现结论总体上和线性回归比较相似。发现演奏、健身、鬼畜调教、音MAD、明星等频道和搞笑、vocaloid、动画等类型能够显著提高视频的播放量，而资讯、星海、机械、手工、广告等频道和说书类型则会显著降低播放量。UP 主的性别影响不高，视频时长的系数则收缩至 0，没有明显影响。

回归树：



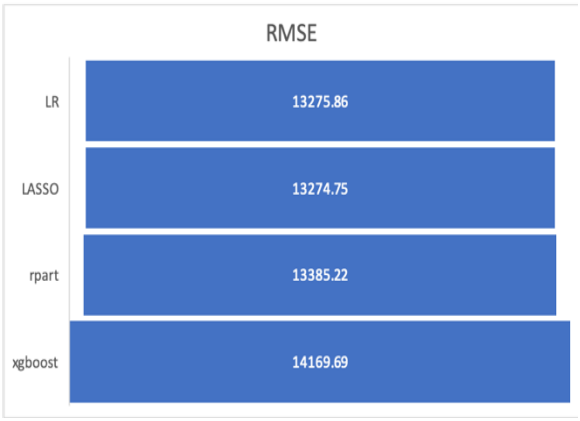
我们发现，该树先从鬼畜频道进行分裂，然后考虑视频时长和是否为搞笑类型。接下去考察专栏和动画类型。

XGBOOST 回归分析：



在 xgboost 模型中，视频时长反而占据了重要地位。然后是是否有专栏。类型、性别和频道也是比较重要的变量。

综上所述，我们发现搞笑、vocaloid、鬼畜、动画等 B 站传统题材有助于提高播放量，同时一些新兴的题材如明星、音乐等也得到用户的喜爱。视频时长对播放量有明显的影响，但可能因为其影响并非线性，因此在线性回归和 LASSO 中不显著，却在树模型中占有重要地位。



回归效果比较：

对以上这 4 个模型[7]做五折交叉验证：

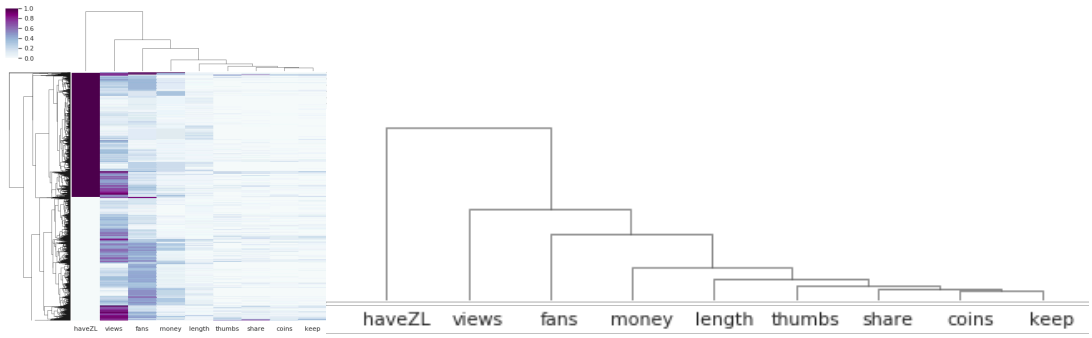
RMSE 最低的为 LASSO，为 13275，然后是线性回归，为 13276，回归树和 xgboost 两个模型效果略差，其 RMSE 分别为 13385 和 14170。

[7] 自变量选取非“收藏，投币，分享，点赞”等事后反馈变量，而是选取可以事前控制的自变量

四、聚类模型

本部分我们也将针对视频标题进行了文本分析，并就 UP 主的情况，如是否有专栏、粉丝数等进行了聚类分析。

在将数据归一化之后，我们发现如下模式：



可以看出，几个变量按照顺序依次形成聚类，收藏，投币，分享，点赞依次汇聚。这四个变量都是视频发出后得到的反馈，视频时长在这里汇入，这个是 UP 主可以事前控制的，其次是充电（打赏），同样是事后反馈，粉丝数是基于积累，播放量在此聚入，说明播放量与前面的复合变量相关性较大。最后聚入的是专栏数。从此结果可以看出想要更高的播放量需要对视频时长多加把控。

最后，我们针对视频标题使用 Latent Dirichlet Allocation 模型进行了文本分析。

吃	0.024	杂谈	0.049		0.049	游戏	0.038	期	0.029	六道	0.101	童年	0.027
笑点	0.019	主播	0.031	直播	0.012	守望	0.016	木鱼	0.017	绅士	0.028	经典	0.02
咬人	0.015	一分钟	0.029	游戏	0.011	先锋	0.014	剧场	0.014	中国	0.015	秀	0.013
鸡	0.014	篇	0.026	「	0.01	VS	0.01	盘点	0.014	散人	0.013	欢乐	0.011
毛蛋	0.013	TOP10	0.023	天天	0.009	恐怖	0.009	翻唱	0.013	解说	0.013	鬼畜	0.01
♥	0.013	玩	0.022	开箱	0.009	小强	0.009	炸	0.012	时刻	0.011	期	0.01
	0.012	小点	0.021	谷歌	0.009	文曰	0.009	微	0.011	视频	0.011	世界	0.008
舞蹈	0.009	真会	0.014	黑桐	0.008	分钟	0.009	小缘	0.011	BOY	0.009	抽风	0.008
做	0.007	绘画	0.008	科技	0.008	作	0.008	站	0.009	日本	0.007	努巴尼	0.007
爱	0.006	老师	0.008	美学	0.007	陆夫人	0.008	B	0.008	新番	0.007	游戏王	0.006
西四	0.006	死	0.007	喵	0.006	读	0.007	龙珠	0.008	小熊	0.007	LOL	0.006
ŏ	0.006	粉丝	0.007	中国	0.006	故事	0.006	C	0.007	集锦	0.007	水彩	0.005
ω	0.006	野食	0.006	手绘	0.006	wanna	0.006	炉石	0.007	方言版	0.007	长歌	0.005
Benny	0.006	青铜	0.006	试玩	0.005	火影忍者	0.006	英雄	0.006	猫	0.006	厨房	0.005
分享	0.006	修炼	0.006	猫	0.005	神奇	0.006	素材库	0.006	四川	0.006	画	0.005
欣小萌	0.005	手册	0.006	新	0.005	原著	0.006	周六	0.006	推荐	0.006	版	0.005
全明星	0.005	徐	0.005	体验	0.004		0.005	局长	0.006	动画	0.006	开心	0.005
番茄	0.004	超神篇	0.004	君	0.004	说	0.005	萧忆情	0.005	游戏	0.006	躺	0.004
黑暗	0.004	鬼	0.004	美	0.004	猫	0.005	撒	0.005	实况	0.006	十大	0.004
第一次	0.004	小哥	0.004	解说	0.004	王老菊教	0.004	痒	0.005	挑战	0.006	污妖王	0.004

LDA 算法文本聚类，将视频标题中的文本抽取关键词分为 7 类：**可爱吃货区，标题党，天天看直播，游戏对战，小众，文化，2B 青年欢乐多**

从模型意义上来说，这反映了 B 站上的视频比较主流的就是这些类别，而其中也有受欢迎程度之分，后者在回归结果中反映。

我们看出在每个聚类中关键词的主题还是相对分散，LDA 在英文文本主题聚类上具有较好的

结果，而在此数据集上效果可能还是一般。如果仔细观察被选出的高频关键词，比如笑点，超神，小哥，开箱，，君，守望先锋等，可以看出这些都是当下青少年口中的流行词，俗称‘黑话’。现在的青少年，看 B 站主要看的也就是游戏（手游、页游、网游、守望先锋、吃鸡等）和鬼畜，再就是动漫讲解（上面有‘火影忍者’、‘新番’ 等关键词）等等，而这类视频在 B 站上有很多，在内容已定，且不考虑 UP 主粉丝效应的影响下，视频时长和标题在直观上是最能控制的自变量。

我们看“改革春风吹满地”这个标题，看似又红又专，实际上是个鬼畜，再看“【LOL】我是人鱼小公举”就感觉是个无聊标题，指向性不明确，自然也没有太多吸睛的效果。

以上给我们的指导意见是，如果要播放量高，在视频标题上肯定也要下功夫，一定要选择吸睛的，针对主流受众（青少年群体）的‘黑话’，内容做不好，至少可以标题党嘻嘻。

五、结论及反思

通过研究，我们发现 B 站现在已经从一个只有 ACG 内容的二次元网站发展成为一个以二次元内容为特色，包含丰富内容的综合性视频网站。纵览所有样本中人气指数较高的视频，我们发现其中不仅仅有 ACG，还有音乐、舞蹈等内容。在内涵不断丰富、用户不断增多的同时，B 站也有更高的能力打造网红偶像。从去年爆火的华农兄弟，到去年末入驻 B 站，现在已有 120 万粉丝的吃播日常 UP 主“徐大 SAO”，我们似乎可以看到 B 站已经能够将那些提供优质视频的 UP 主捧成新的网红偶像。

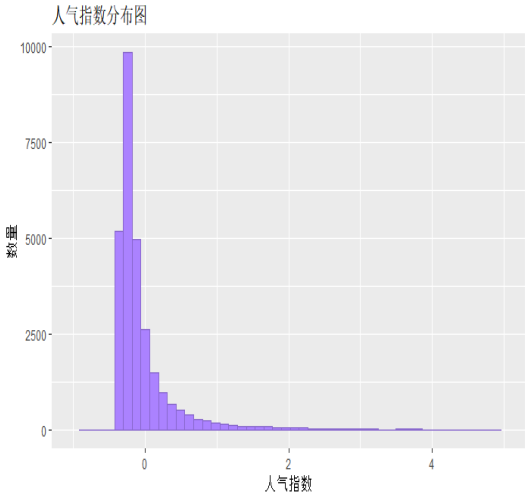
但同时也应当注意到 B 站 UP 主的竞争是非常激烈的。有月均 20 万新粉丝的 UP，也有更多苦苦经营几年却最终只有几千粉丝的 UP。因此，我们希望结合我们建模的结果，谈谈一个现实的话题：如何增长人气，即对于新进入 B 站大家庭的小白 up 主，如何才能马上变网红，走上人生巅峰？

首先，我们认为选择自己的**up 主定位**很重要，虽然现在 B 站内容逐渐多元化，但我们还是发现游戏和鬼畜这两个传统阵地具有更多受众。因此，选择做一个游戏 up 或者鬼畜 up 虽然竞争更加激烈，但也更容易走红。其次，**视频的时长**也值得注意。虽然长视频可能更有内涵，但通常而言用户更青睐短视频，因此用短视频起步不失为一个好策略。最后，如果看到自己精心制作的的视频播放量上不去，而一些知名 up 随便糊弄的视频都播放量爆表也不要过于沮丧，这可能是因为他们的粉丝基数大，因而被推荐的几率也会更高。视频中也千万不要怕尴尬，勇敢向观众**索要三连**吧！

六、附录——一些有趣的工作

附录 1、分区热门视频解析

我们希望根据以上五个指标对视频热度进行解析。首先，我们对以上 5 个变量进行主成分分析。我们发现，前两个主成分能够解释的方差占比为 0.78 和 0.11，因此我们采用这两个主成分进行讨论。我们计算了各个样本的主成分 1 和主成分 2 的值，并按照其解释的方差的比例进行加权，其中主成分 1 占比为 0.78/（0.78+0.11）=0.88，主成分 2 占比为 0.12。我们称加权后的结果为该视频的“人气指数”，并希望探讨不同分区中人气指数最高的是哪些视频。



左侧为人气指数的分布图。我们可以发现大部分的人气指数分布在-1 到 2 之间。其最小值为-0.35，最大值为 62，中位数为-0.20.

视频人气指数最低：《L 星人 snake 经理左雾：就算 0.1%的帮助我也会去做》

视频人气指数最高：《改革春风吹满地》

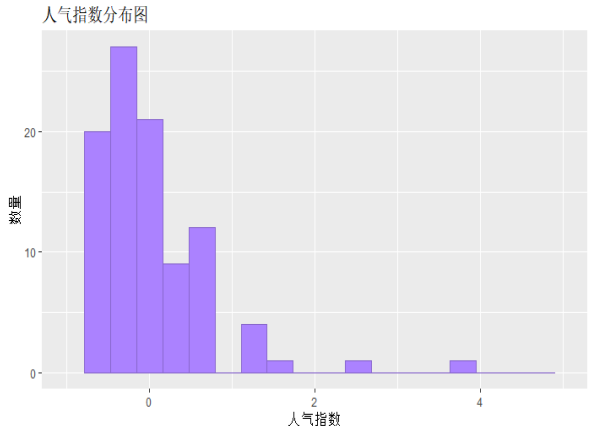
B 站中的视频有不同的频道。我们挑选了样本中视频数量大于 500 的频道作为热门频道（一共 13 个这样的频道），并从每个频道中挑选出 5 个人气指数最高的视频，共 65 个。不同 UP 主上榜视频数量表如下：

up 主	热门视频数	备注
爱做饭的芋头 SAMA	5	美食圈频道
LexBurner	5	5 个均为综合频道，但内容为动画点评
科技美学	5	专注于手机平板评测
机智的党妹	5	美妆
芒果冰 OL	5	电子竞技、网游、手游
信誓蛋蛋	4	电子竞技、日常
老番茄	4	专注于单机游戏解说
敖厂长	4	单机游戏
神奇的老皮	3	4 个电子竞技频道，2 个搞笑频道
伊丽莎白鼠	3	鬼畜调教

每个分区人气指数最高的视频分别为：

视频名称	分区	UP 名称
烂尾的游戏冒险(雅达利寻剑)	单机游戏	敖厂长
暗影崛起-最秀新卡发布！	电子竞技	神奇的老皮
【不要怕辣挑战】试吃 5 种最辣食物！！	搞笑	徐大虾咯
对于做成啵啵茶的芋头，真是毫无抵抗力啊……	美食圈	爱做饭的芋头 SAMA
《复联 4》前必看！一口气看完 21 部漫威电影，完整的时间线剧情讲解！	影视杂谈	努力的 Lorre
网红挑战荒岛求生用 iPhone 成功抓鱼【第七集】	日常	信誓蛋蛋
剧情刺激！场面爆笑！这谁顶得住啊！2019 一月新番吐槽大盘点！	综合	LexBurner
英文版《改革春风吹满地》！！Chinese people so 牛逼！	翻唱	A 路人
【中文八级】两个国人展开了惊人的英语八级对话	网络游戏	某幻君
【春晚鬼畜】赵本山：我就是念诗之王！【改革春风吹满地】	鬼畜调教	小可儿
【党妹】改造闺蜜|第一次体验 Lolita 和拍写真！蔷薇色温柔少女妆+和风浴衣发型，从日常到上镜只需一点心机	美妆	机智的党妹
【游戏侦查冰】揭秘游戏氪金抽卡的“黑幕”	手机游戏	芒果冰 OL
「科技美学直播」拼多多爱爱"旗舰手机体验，那岩已经疯了"	手机平板	科技美学

同样，我们也计算了 UP 人气指数。该指数考虑了视频人气指数、粉丝数和充电数。由于充电数有缺失值，因此我们没有将充电数为 NA 的 up 纳入考虑中。该指数使用主成分分析，考虑视频人气指数、粉丝数和充电数。因为每位 UP 都发布多个人气指数不同的视频，因此每位 UP 会得到多个结果。我们将 UP 人气指数设定为这些结果的算术平均。以下右边是人气指数最高的 6 位 UP，左侧是人气指数分布图。



up 主	up 主人气指数
敖厂长	3.77
LexBurner	2.51
老番茄	1.49
伊丽莎白鼠	1.41
渗透之 C 君	1.39
papi 酱	1.26

根据人气指数和分区热门我们可以模拟出 B 站视频晋升机制排名，从而有针对性地设计和改良自己的视频

附录 2 老番茄简介

关注数	粉丝数	播放数
1	461.3万	3.8亿

“你可以在老番茄身上看到自己的影子。老番茄就是我们的同龄人，你可以看见他的成长和精进，不管是学业上的、生活上的还是视频上的。他代表你实现了人生的一部分可能性。”

复旦金融系大四学生

著名 B 站 UP 主

知名网络播客和游戏娱乐解说人

粉丝数：367 万

自制视频总播放量：2.9 亿

成名作：《你可见过如此丧心病狂的口袋妖怪解说》系列。

[i] 极光大数据 2017 年统计报告