



清华大学
Tsinghua SEM



商务分析整合实践期末报告

快手：用户行为预测 —— 取消关注

小组成员：
宋辰菲 2018211650
肖念瑶 2018211660
闫嘉文 2018211671

指导老师：郭凡、廖闯

12/26/2018



清华经管学院
Tsinghua SEM



研究问题

1. 分析海量用户数据，了解取消关注的机制；
2. 建模预测取消关注事件的发生；

问题意义

研究用户取关行为是有意义、有必要的

- 取关行为可以分成两类：合理 / 非合理取关
- 避免、减少非合理取关行为
- 允许、引导合理取关行为
- 需要我们分别细化研究，引导社区建设



01

I. 快手公司



APP (14)

好友数量 – follows
内容分发机制 – shows

02

II. 消费者



关注强度 (7)

关注路径
相互关注
平级关注

03

III. 生产者



内容整体质量 (14)

平均作品点击数
平均作品点赞数
平均作品评论数
播放时间

04

IV. 互动行为



A-B 互动行为 (6)

展示数
点赞数
评论数
播放时间





目录

Content

01

变量准备

02

趋势分析与案例研究

03

预测模型

04

总结





/01

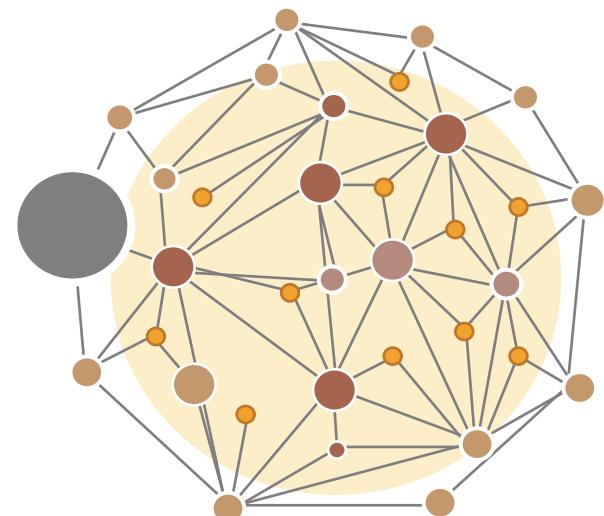
数据准备

1. 已有数据
2. 变量构建

一、数据准备

1. 汇总数据

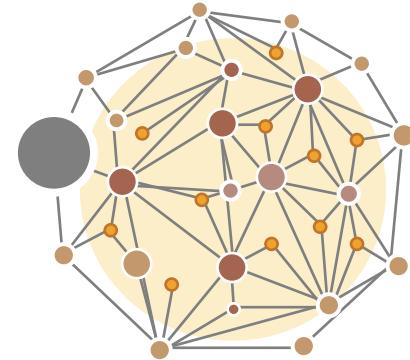
变量	含义	形式	变量类型
from_uid	发起关注用户ID	Egwyfgw (小王)	字符串
to_uid	被关注用户ID	Hsdhjbs (隔壁老王)	字符串
follow_time	关注时间戳(毫秒)	1535776508000 (9.1, 12:35)	时间戳
follow_stack_array	关注栈(关注路径)	[2] (用户搜索)	列表
is_unfollow	是否取消关注	1	0-1变量
unfollow_time	取关时间	1538311380000 (9.30, 20:43)	时间戳
unfollow_page_ref	取关路径	[4] (好友列表)	列表



一、数据准备

2. 社交网络数据

变量	含义	形式	变量类型
From_uid	发起关注用户ID	Egwyfgw 小王	字符串
To_uid	被关注用户ID	Qy3gsads 楼上小芳	字符串
Follow_time	关注时间戳(毫秒)	1531573980 (7.14)	时间戳
Follow_stick	关注栈(关注路径)	[26] (扫一扫)	列表



3. 用户关注后的视频消费数据

变量	含义	形式	变量类型
From_uid	发起关注用户ID	Egwyfgw (小王)	字符串
To_uid	被关注用户ID	Hsdhjbs (隔壁老王)	字符串
Sever_time	视频展示时间	1537274580000 (9.18)	时间戳
Photo_id	视频ID	8164265919	整型
Is_click	是否播放	0	0-1变量
Playing_times	播放时长(毫秒)	NULL	整型
Is_like	是否点赞	0	0-1变量
Comment_times	评论次数	0	整型

一、数据准备

4. 用户关注后的固有属性特征

变量	含义	形式	变量类型
User_id	用户ID	Hsdhjbs (隔壁老王)	字符串
active_data_list_30d	过去30天活跃日期列表	["20180908","20180915","20180922","20180904"]	列表
register_time	注册时间	2011/10/1 10:02	日期
fre_province	过去30天常驻省份	北京	字符型
fre_city	过去30天常驻城市	北京	字符型
fans_cnt	粉丝数	1798402	整型
follow_cnt	关注数	385	整型
friend_cnt	好友数	189	整型
photo_id	过去30天内上传作品ID	NULL	整型
live_stream_id	过去30天内打开直播ID	3450804815	整型
photo_or_live_create_time	作品或直播创建时间	2018/9/15 18:56	日期
photo_or_live_province_name	作品或直播创建时所在省份	北京	字符型

一、数据准备

变量选择与构建

01

I. 快手公司



APP (14)

好友数量 – follows
内容分发机制 – shows

02

II. 消费者



关注强度 (7)

关注路径
相互关注
平级关注

03

III. 生产者



内容整体质量 (14)

平均作品点击数
平均作品点赞数
平均作品评论数
播放时间

04

IV. 互动行为



A-B 互动行为 (6)

展示数
点赞数
评论数
播放时间

一、数据准备

“宽” 数据构建结果

41个研究变量

220万关注与取消记录

From_uid	To_uid	Follow From	Fan_from	Reci Follow	Follow Stack	Click To	Clicks	...
A	X	43	120	1	直播	10	1	...
A	Y	43	120	0	搜索	17	0	...
B	X	21	68	0	同城	10	0	...
C	Y	52	10	1	直播	17	2	...
C	Z	52	10	0	搜索	5	1	...
...

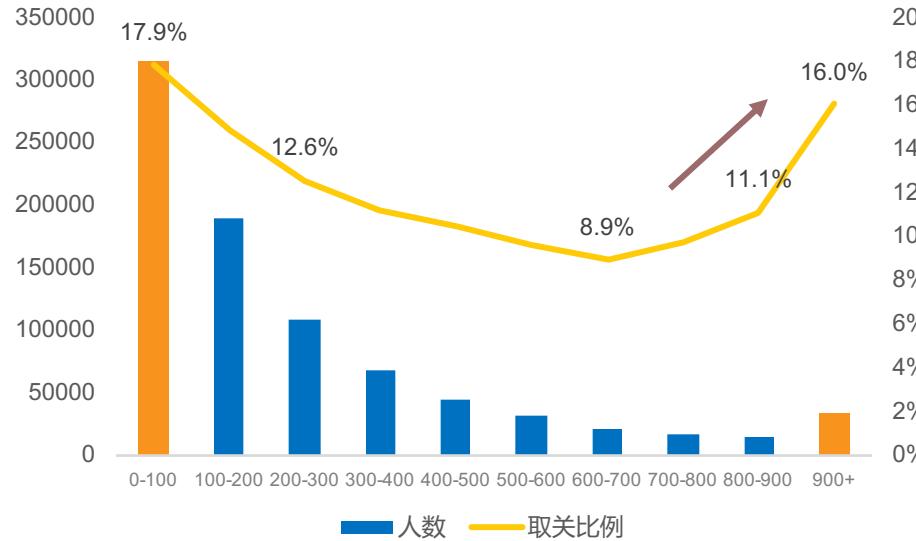


/02

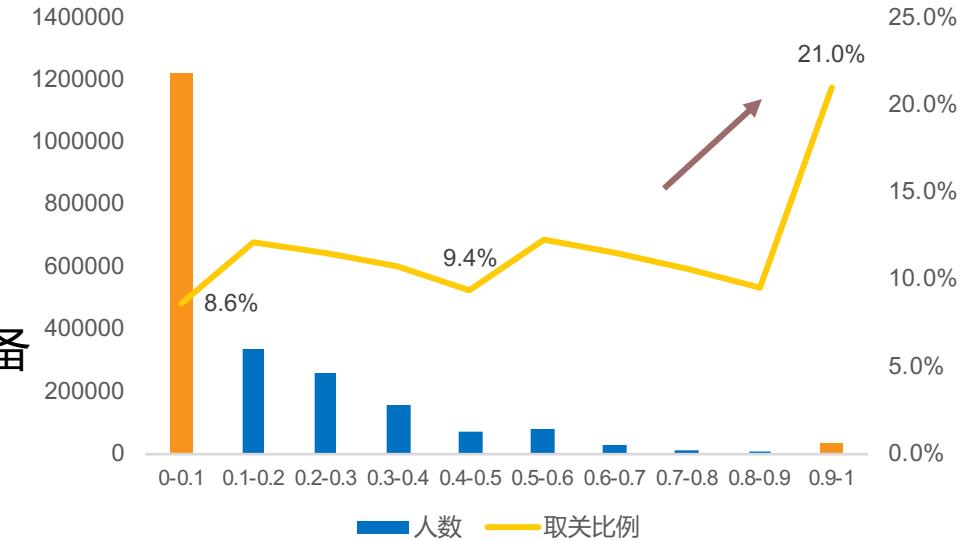
趋势分析与个例研究

1. 单变量趋势分析
2. 案例研究

二、趋势分析 - APP机制 & 取关率



一、数据准备



变量 : follow_from

发现 : 关注人数 & 取关率

- ✓ 关注人数<100 : 取关率17.9%
- ✓ 关注人数>700 : 取关率逐渐上升
- ✓ 关注人数>900 : 取关率16.0%

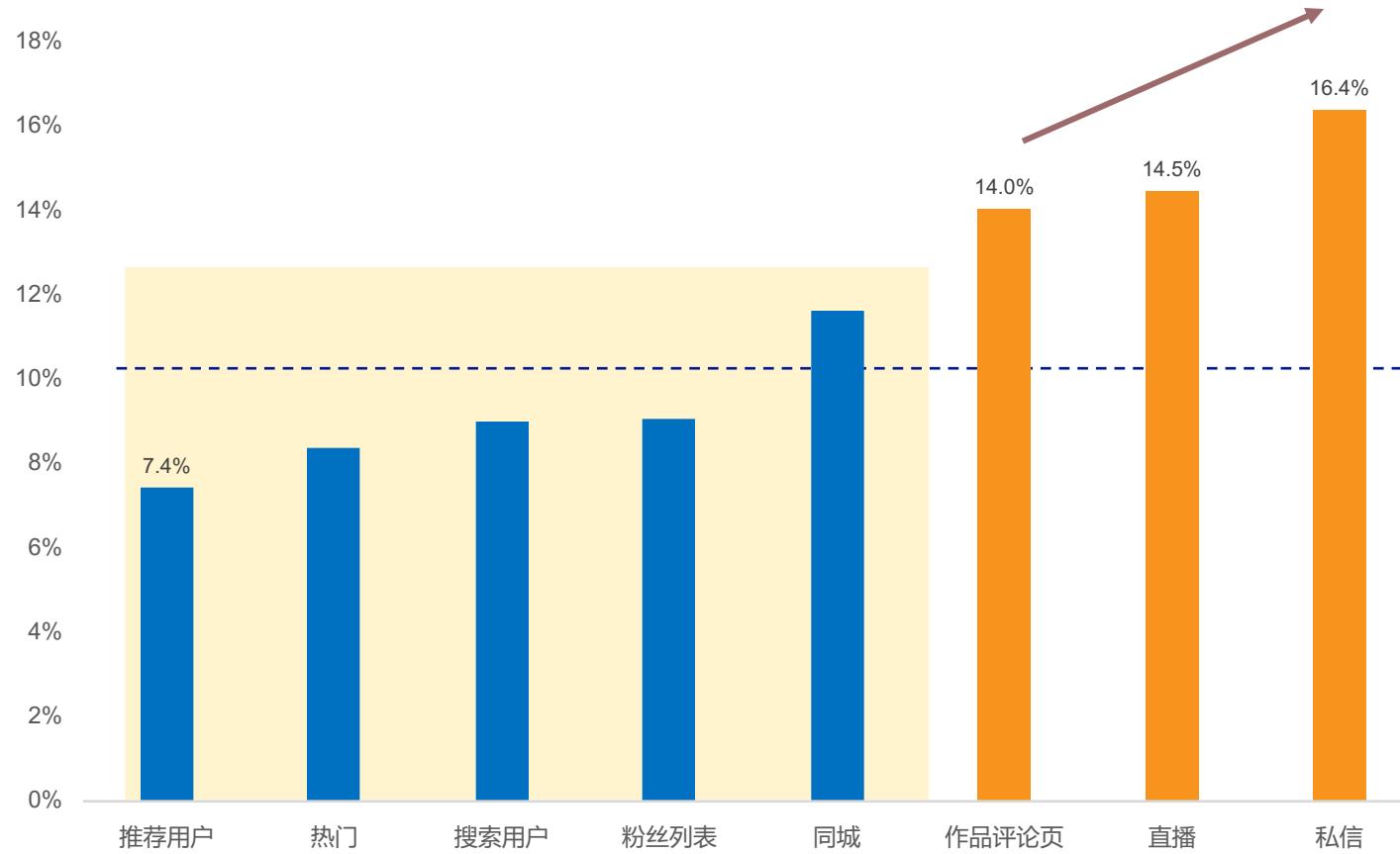


变量 : click / show

发现 : 点击率 & 取关率

- ✓ 点击率<0.1 : 人数占比50%以上
- ✓ 点击率>0.9 : 取关率21.0%

二、趋势分析 - 用户本身 & 取关率



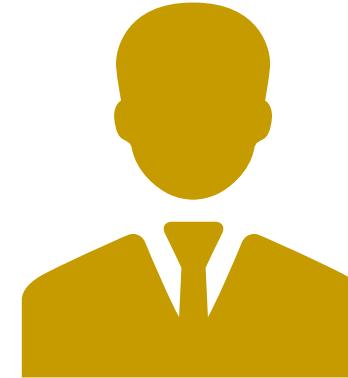
- 变量 : follow_stack
- 发现 : 关注路径 & 取关率
- ✓ 通过作品评论页/直播/私信关注
取关率较高
 - ✓ 推荐用户取关率最低

二、趋势分析 - 用户本身 & 取关率



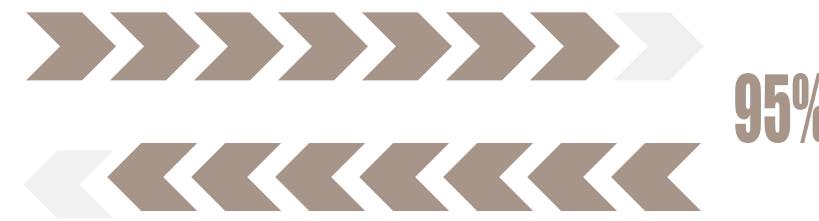
单向关系 (One-way)

更容易破裂: 2,003,160人中继续关注率为89.4%



双向关系 (Reciprocal)

更不容易破裂: 210,779人中继续关注率为94.98%



二、趋势分析 - 用户本身 & 取关率



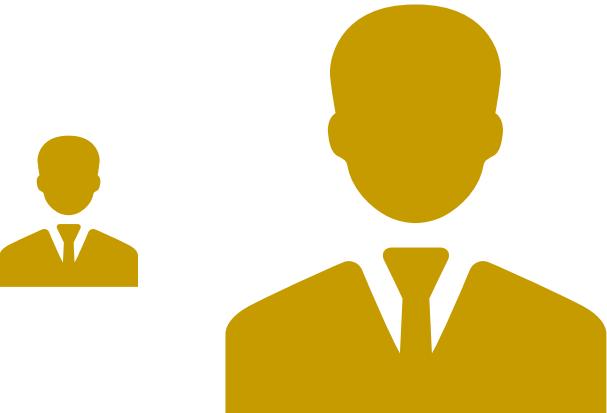
平等关系 (Equal)

更容易破裂: 121,302人中继续关注率为85.9%



非平等关系 (Unequal)

更不容易破裂: 2,092,637人中继续关注率为90.2%



平等 = A , B两人粉丝数量比 [0.5, 2]

(稳健性检验 : 比值为[0.2, 5]或[0.1, 10]的结果类似)

二、个例研究：APP关注上限与密集取关



ID : bOK5MWliKTJLAWc4eEkV8Q==

998
关注数

95%
取关率

95%
取关发生在两分钟内

2018年9月1日



关注20人

2018年9月2日

07:41
取关11人



07:47
取关1人



07:42
取关7人

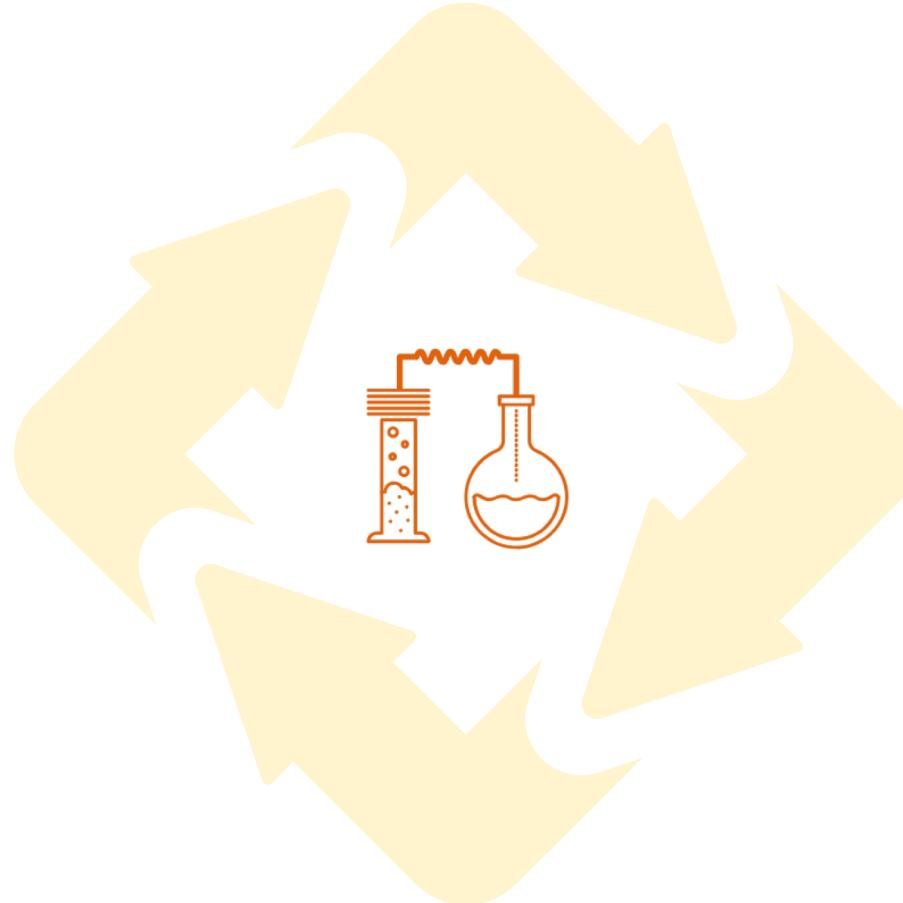
二、个例研究：实验设计

软门槛

用户关注达到900、950、990时发出提示

预测结果

在900、950、990时会有大批比较密集的取关事件发生



硬门槛

给部分测试用户提高关注人数权限

预测结果

在990附近不会出现密集取消关注事件的发生

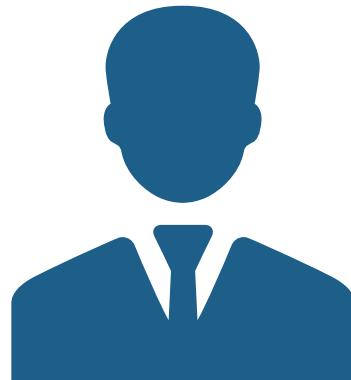
二、个例研究：骗粉



如何识别骗粉？

- 双方有互相关注，且互相关注的时差不会太长
- 关注之后双方并不会产生互动，关注者不会点击、观看、点赞、评论被关注者的作品
- 关注之后不长时间内便会发生取消行为
- 关注途径极有可能是通过私信【6】、消息【17】、直播【7】、用户搜索【2】等，热门【8】、用户推荐【1、18、19、20、21、22】则比较少

二、个例研究：骗粉



粉丝数 : 943

ID : dmALn6teQKC5
np+x1nZXmQ==

发起关注 : 2018/9/1 8:34 AM 途径 : 私信

互相关注 : 2018/9/1 8:34 AM

取消关注 : 2018/9/3 1:37 AM

0点击，0点赞，0观看



粉丝数 : 204

ID : 0p9a8a6W/TXpC
Z9ecKVEsg==



/03

预测模型

1. Logistic Regression
2. 模型比较、特征重要性、聚合模型
3. Spark 平台

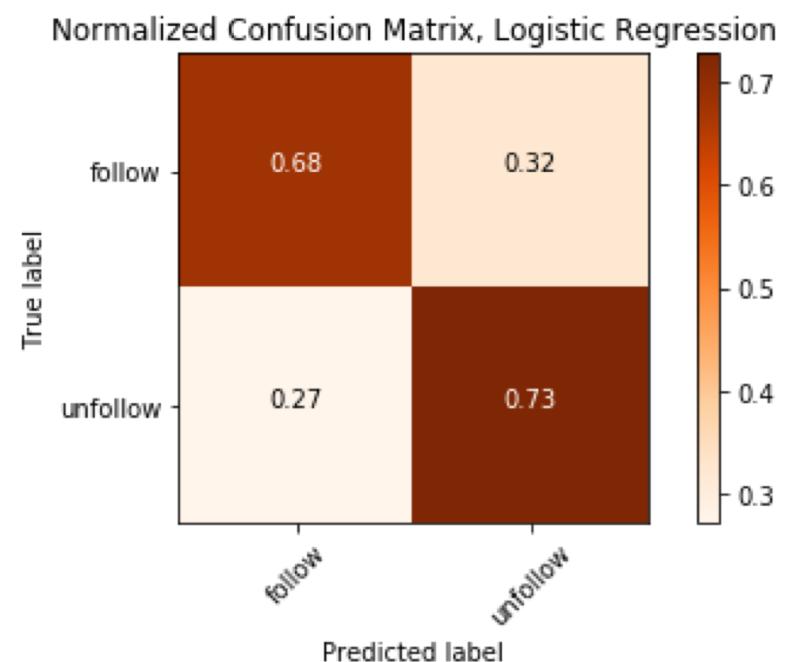
三、预测模型

- **Y : 自变量 is_unfollow**
 - 用户 f 在9月1日关注 u 后，在9月30日继续关注(0) 还是取消关注 (1)
- **X : 因变量**
 - APP : 7 个；关注强度 : 7 个；内容质量 : 14 个；互动行为 : 6 个 (构建哑变量 2 个)
- **预测模型**

- **Benchmark Model**

- **Logistic Regression**

Full	Logistic Regression
In-sample	70.1%
TN	72.8%
Precision	68.0%
Recall	71.4%
Accuracy	70.4%



三、预测模型

- 更多分类预测模型

Full	Logistic Regression	Linear SVC	Decision Tree	SGDC	Naïve Bayes	NN1 (256,)	NN1 (128,)	NN1 (64,)	NN1 (32,)	NN2 (128,64,)	NN3 (128,64,32,)
In-sample	70.1%	69.6%	62.0%	67.1%	59.6%	74.5%	74.2%	73.9%	73.4%	73.1%	73.8%
TP	68.0%	67.0%	63.4%	78.4%	51.2%	78.8%	78.1%	79.2%	79.4%	78.4%	79.2%
TN	72.8%	73.5%	62.8%	55.5%	66.5%	70.4%	69.8%	68.5%	66.6%	68.2%	67.5%
Precision	68.0%	67.0%	63.4%	78.4%	51.2%	78.8%	78.1%	79.2%	79.4%	78.4%	79.2%
Recall	71.4%	71.6%	63.0%	63.8%	60.5%	72.7%	72.1%	71.5%	70.4%	71.1%	70.9%
Accuracy	70.4%	70.2%	63.1%	66.9%	58.9%	74.6%	73.9%	73.8%	73.0%	73.3%	73.4%

- LinearSVC TN (True Negative) 预测准确率最高 (73.5%)
- Neural Network (单隐层 256个神经元) 整体预测准确率最高 (74.6%)

三、预测模型

- 分组建模比较预测准确率

- 分组对象

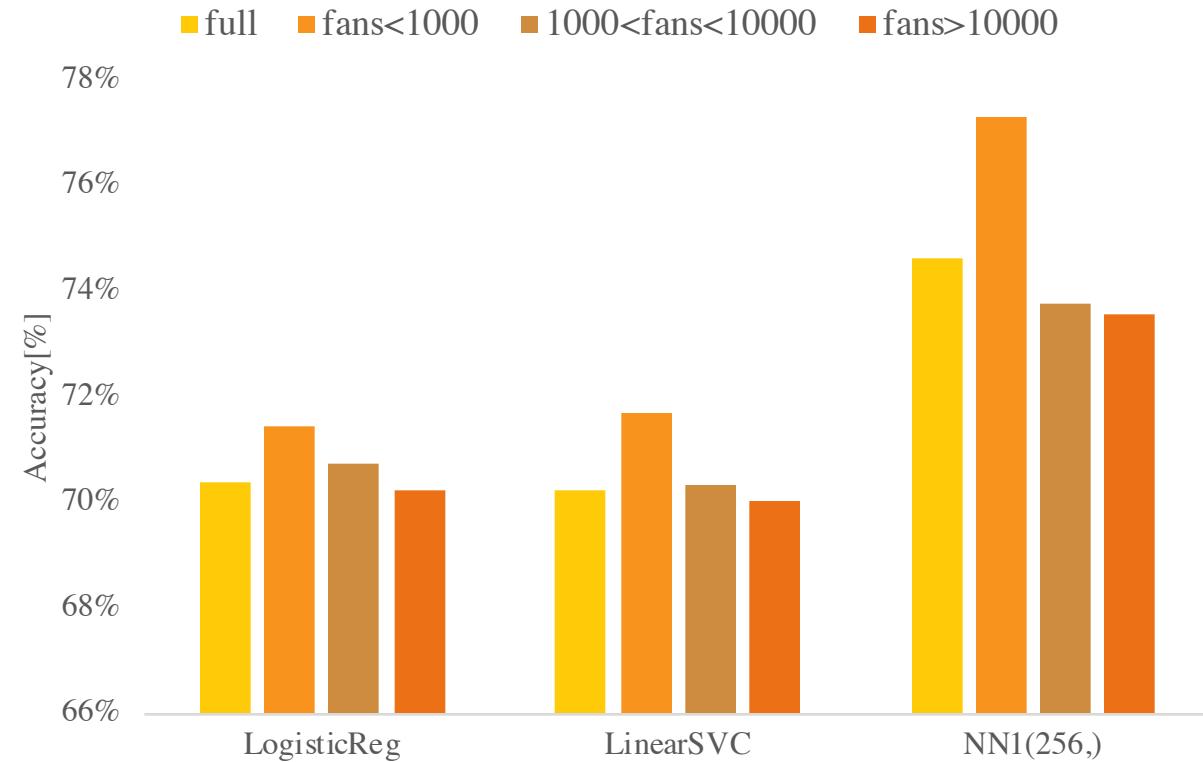
- 粉丝数

- 分组标准

- 以1000和1000为界将数据分为三组

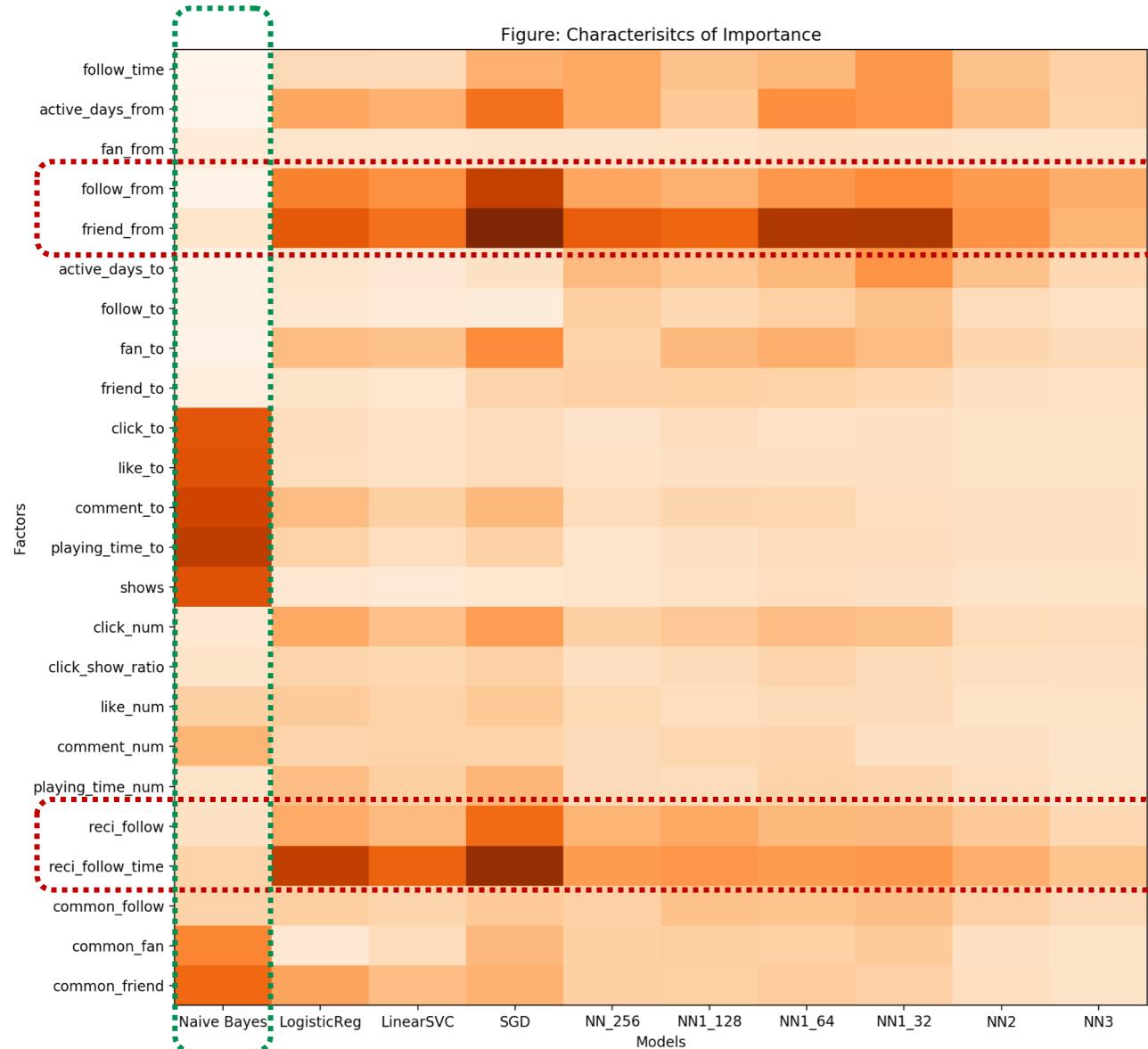
- 预测结果

- 预测准确率整体较高
 - 对普通用户(fans<1000)准确率有明显提升



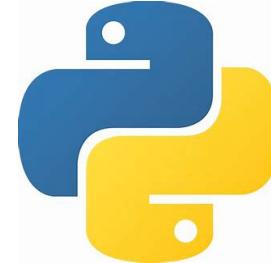
三、预测模型 - 特征重要性

- Factor of Importance Plot
 - 用颜色深浅代表特征的重要性
 - 横向比较：
 - 同一变量在不同模型中的重要性
 - “共同关注”、“共同好友数量”
 - 纵向比较：
 - 不同变量在同一模型中的重要性
 - Naïve Bayes
- Ensemble模型
 - 效果：74.6% VS 72.3%
 - 分析：模型关注的特征较为相似



三、Spark平台初探

- 使用**220万全量数据**，构建预测模型？
- 快手解决方案：Python/R + Hue + Spark
- 大数据通用计算平台



HUE

APACHE
Spark™

```
1 import pyspark
2
3 spark = pyspark.sql.SparkSession \
4     .builder \
5     .appName('Jiawen') \
6     .enableHiveSupport() \
7     .getOrCreate()
8
9 df = spark.read.parquet("://filepath")
10 # df.sql("select * from Table")
11
12
```



三、Spark平台初探

- 构建45万均衡样本数据集 (Pos 22.2 : Neg 22.3)

- Logistic Regression With LBFGS

- 模型准确率 : 60.5%

- SVM With SGD

- 模型准确率 : 64.2%

(222623, 223016)

LogisticRegressionWithLBFGS Training Error = 0.604874613723

SVMWithSGD Training Error = 0.641877359689

- 准确率没有显著提升

- 并行训练模型有限、RDD数据格式转化、惰性操作



/04

总结与展望

1. 过程总结
2. 未来展望

四、总结与展望

课题总结 – 项目



取消关注背后的机制



取关率建模预测



Spark平台大数据初步实践



引导互动



优化APP机制



加强对骗粉的识别

四、总结与展望

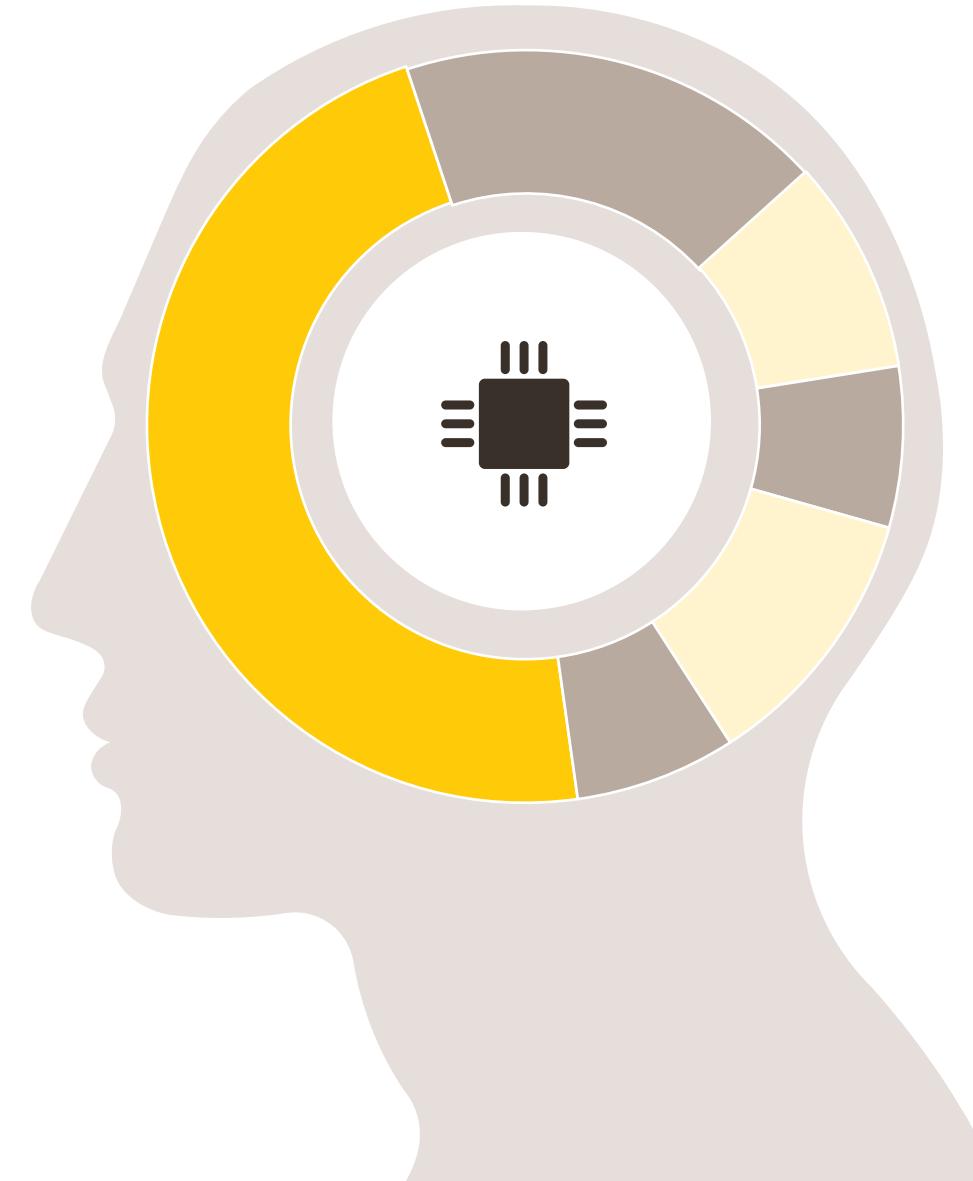
课题总结 – 小组成员

夯实硬技能

- SQL 数据库操作
- Python 机器学习
- Spark 平台

提升软实力

- 小组合作
- 互联网思维方式



四、总结与展望

课题未来展望

- 利用更多数据，对某些特定人群进行追踪研究
- 对用户作品内容、标签等进行探索
- 对模型进行调参，构建更加精确的模型



四、总结与展望

快手项目一瞥





初入只觉伙食好，问题挑战可以秒。

数据规模真不小，细节处理不能草。

团队合作效率高，良师指导不能少。

谢谢！