CSDN×易观流量预测

队伍：橘冯jin

目录

CONTENTS

01 队伍介绍

- 队伍介绍

队伍介绍

队伍名：橘冯jin

队伍成员(昵称)：YWB, 冯波, jin

队伍介绍：在数据挖掘领域,经验丰富，分析角度独特。

02 解读与分析

- 题目解读
- 数据分析
- 难点分析

2.1题目解读

题目解读：

基于某平台海量的真实数据流量历史去预测未来一段时间的流量情况,目的在于了解未来流量趋势可以更快速的调整战略方向，避免不必要的损失，及时抓住相应爆发机会。

预测指标：

pageview:页面浏览量(PV触发次数与UV触发人数)

reg_input_success：成功注册量(PV触发次数与UV触发人数)

2.1题目解读-评估指标

评分方式

预测任务：事件“$pageview“和“reg_input_success”在2019年5月20日至5月26日的PV和UV

$$\sigma_i = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\left(FC_{i,t} - T_{i,t}\right) / \left(T_{i,t}\right)\right)^2}$$

$$F = \sum_{i \in Target} \sigma_i$$

说明：FC是预测值，T是真实值； i 是要预测的目标，t是时间

2.2数据分析

a.kpi_train训练集表基本数据有400条数据，属于小数据

(1) kpi_train训练集数据时间范围2018.11.01-2019.5.19,

即训练集数据是预测时间段的6个半月

b.event_detail训练集表基本数据有2855816条数据，属于大数据

(1) event_detail训练集数据时间范围2018.11.01-2019.5.19,

即数据中title体现中用户的行为画像

c.user_detail训练集表基本数据有569695条数据，属于大数据

(1) user_detail训练集数据时间范围2018.11.01-2019.5.19,

是event_detail行为画像细分用户群体的重要依据

2.3难点分析

难点1：数据的变化有长时间规律也有段时间的波动规律，如何结合

难点2：指标pageview 和 指标reg_input_success是否是独立的

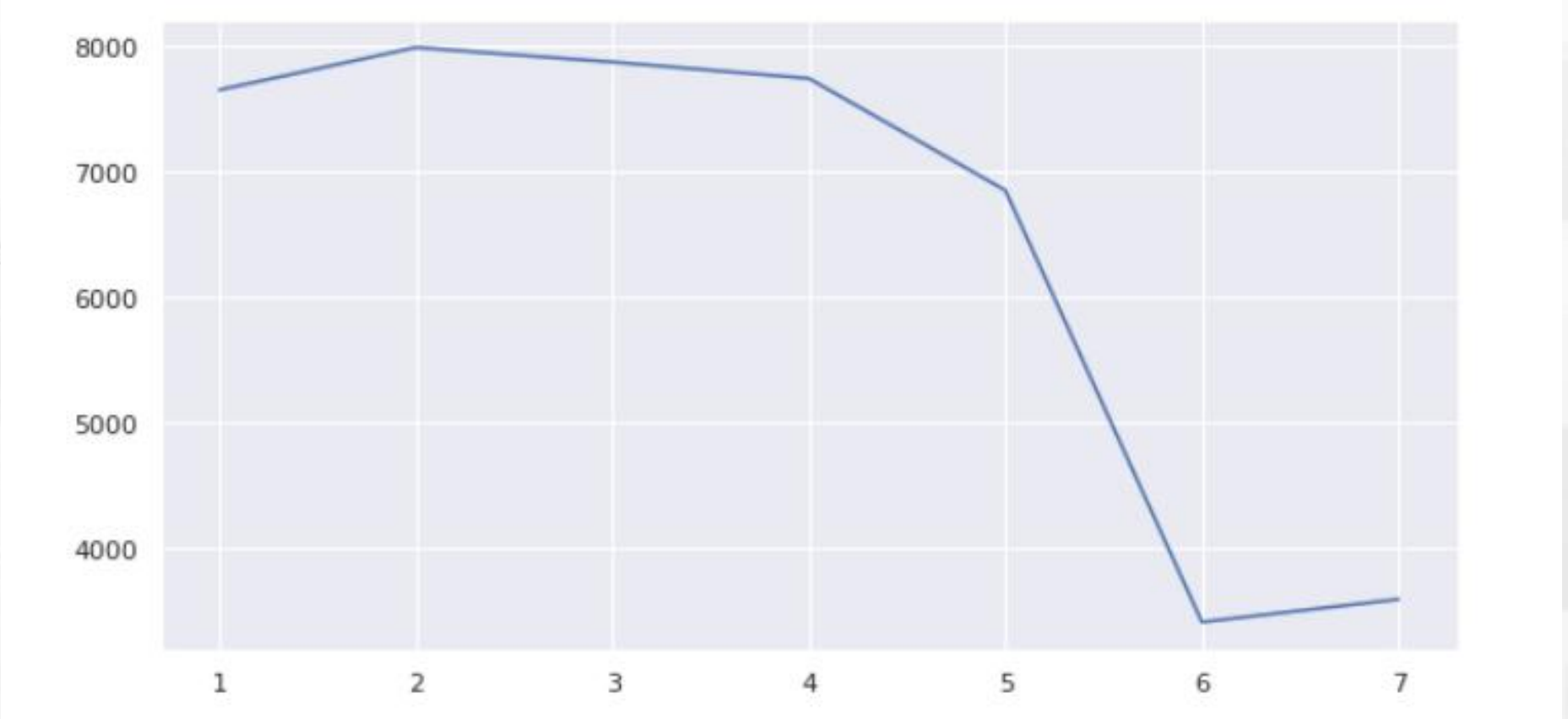
难点3：even_detail与user_detail里蕴含着变化的具体原因，两个预测指标如何跟这两个表关联起来。从中又能挖掘出什么不一样的信息。

03 特征工程与模型

- 行为、流量分析
- 特征工程
- 模型展示

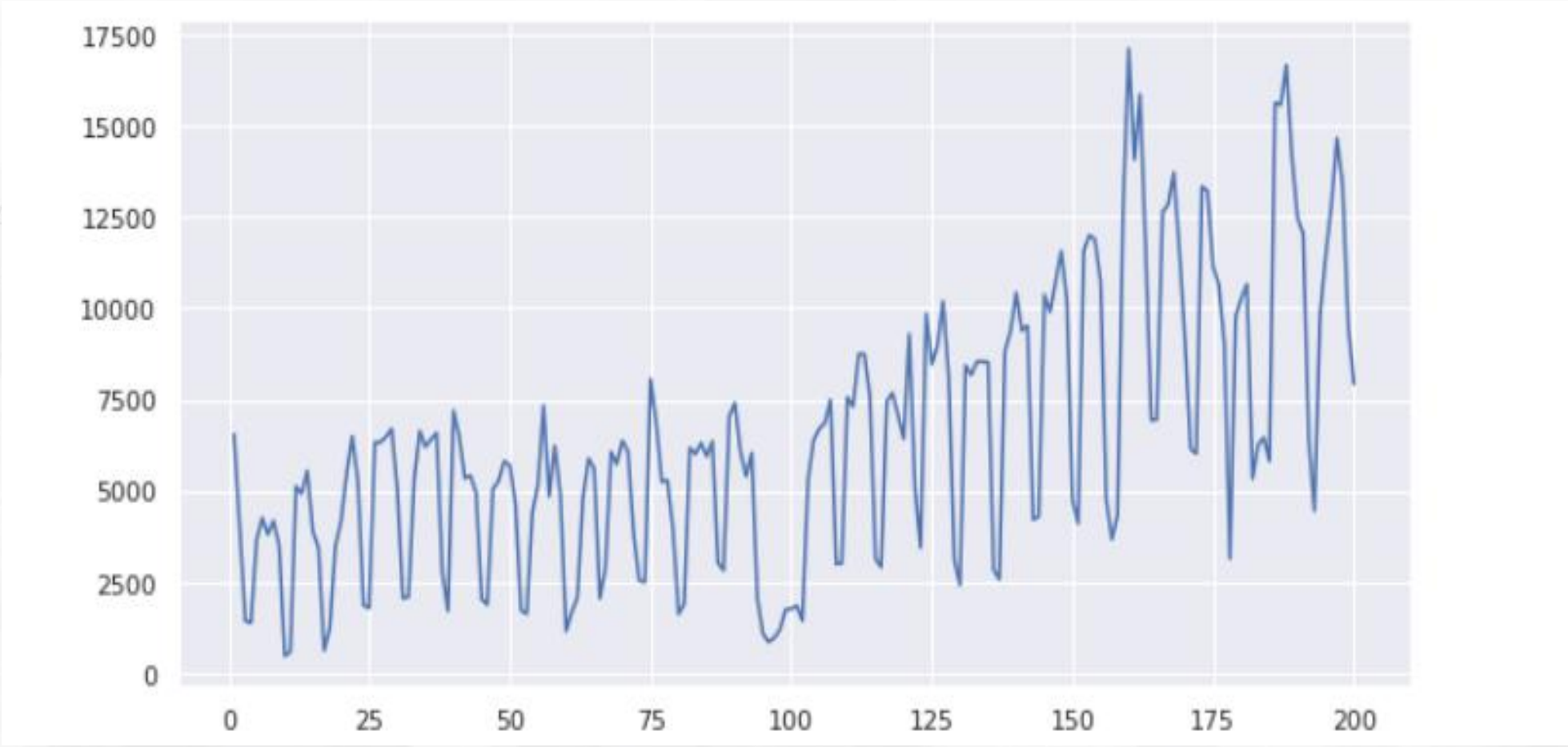
3.1一周规律图

周一到周日 pageview-pv指标



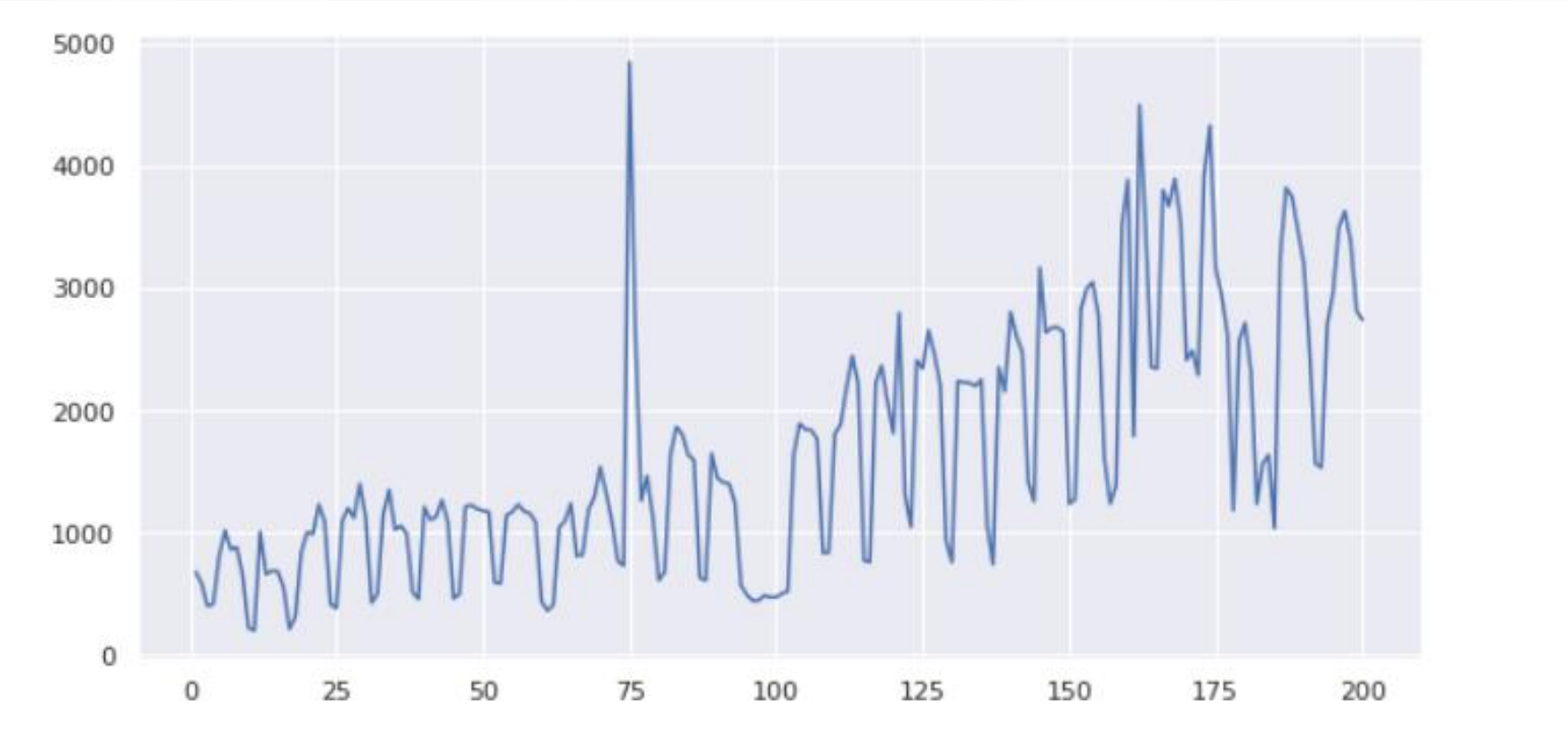
结论：一周内 周六与周日的浏览量最少,符合休息日的休息的规律

3.2历史浏览量pv指标



pageview-pv指标(200天)

3.2历史浏览量UV指标



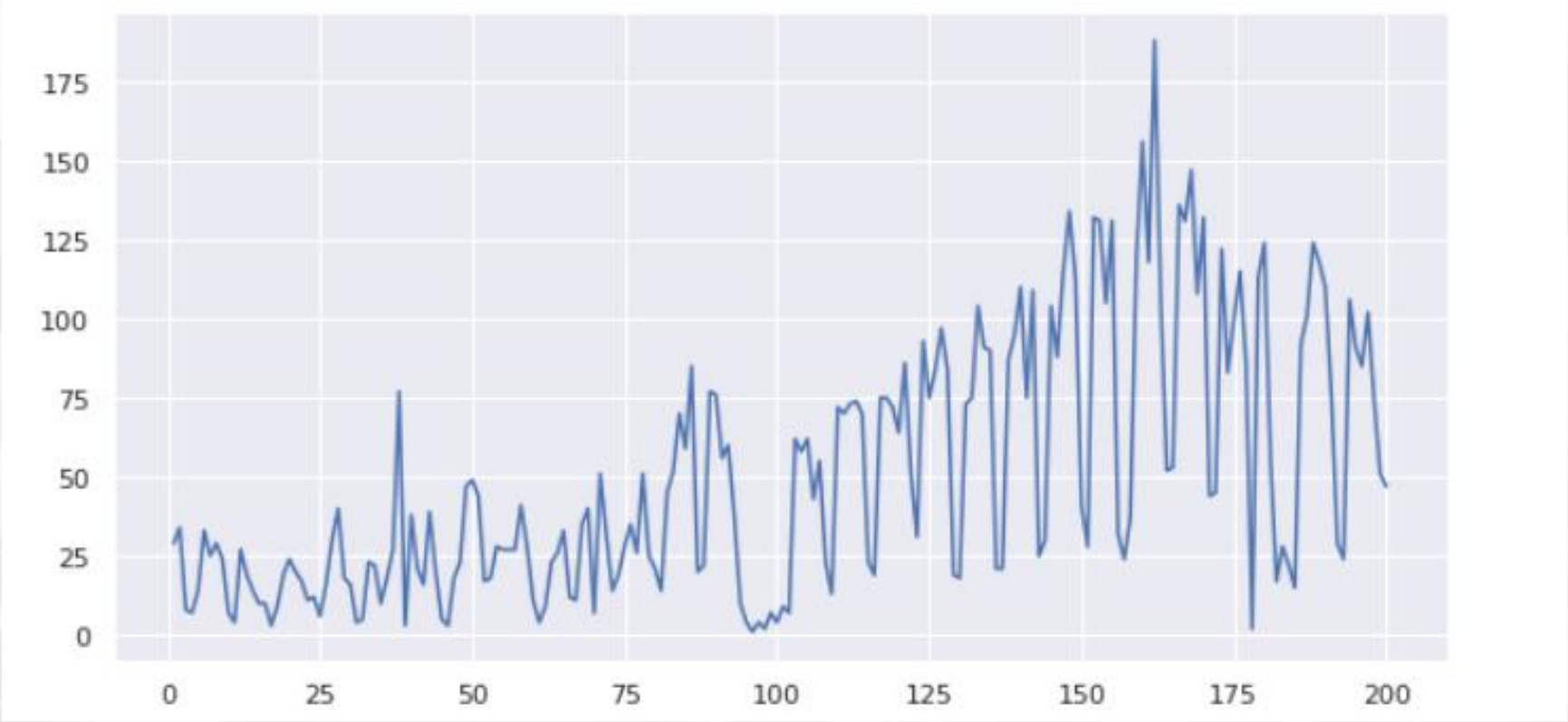
结论：随着时间的增长,平台的企业客户越多，

但最近增势放缓。

3.3浏览量特征

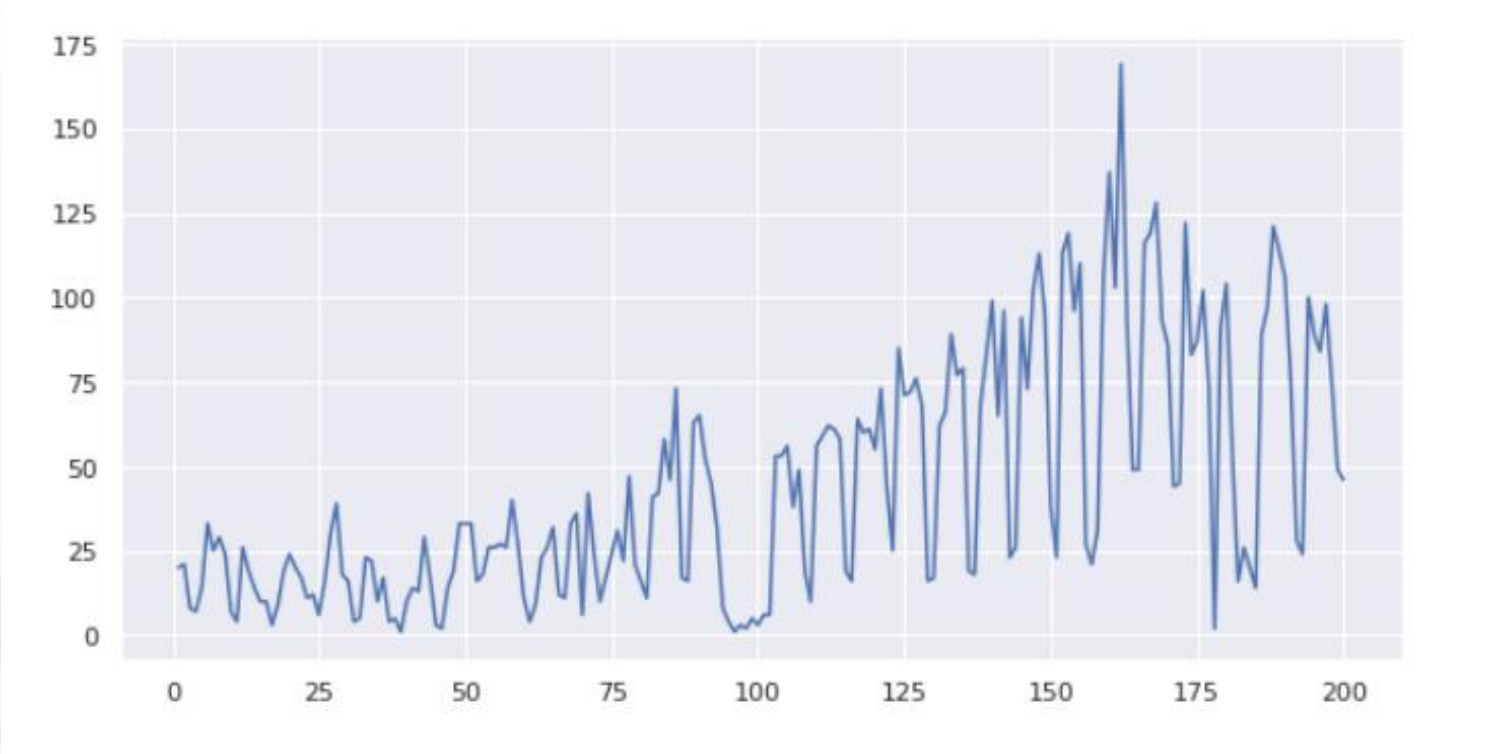
长时间段规律特征	短时间段规律波动特征
2个月内一星期之前的流量量总和	最近两星期的流量均值
2个月内一星期之前的流量量均值	最近两星期的流量方差
2个月内一星期之前的流量量方差	最近两星期同一类别日的流量最大值
2个月内10天之前的流量量总和	最近两星期同一类别日的流量最小值
2个月内10天之前的流量量均值	最近两星期同一类别日的流量均值
2个月内10天之前的流量量方差	当天要预测是一周中哪一天
2个月内两星期之前的流量量总和	
2个月内两星期之前的流量量均值	
2个月内两星期之前的流量量方差	

3.3新注册量PV指标



reg_input_success-pv指标(200天)

3.3新注册量UV指标



结论：随着时间的增长,平台的新企业客户越多,

但近期新企业客户增量有下滑现象。

3.4注册量特征

长时间段规律特征	短时间段规律波动特征
2个月内一星期之前的注册量总和	最近两星期的注册量均值
2个月内一星期之前的注册量均值	最近两星期的注册量方差
2个月内一星期之前的注册量方差	最近两星期同一类别日的注册量最大值
2个月内10天之前的注册量总和	最近两星期同一类别日的注册量最小值
2个月内10天之前的注册量均值	最近两星期同一类别日的注册量均值
2个月内10天之前的注册量方差	当天要预测是一周中哪一天
2个月内两星期之前的注册量总和	
2个月内两星期之前的注册量均值	
2个月内两星期之前的注册量方差	

3.5浏览量与注册量的相关性思考

指标pageview：指的是当天的浏览量

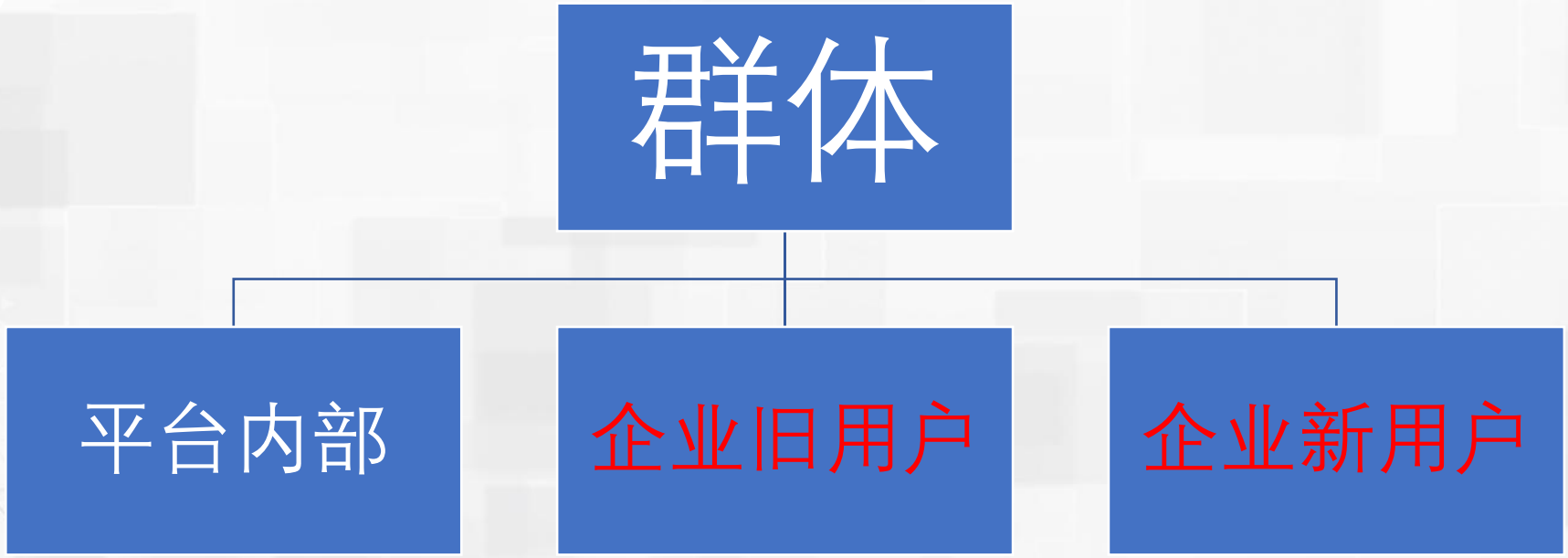
指标reg_input_success：指的是成功的注册量

推论：注册成功意味着有可能转成企业客户，也就有可能转成未来的平台产品使用者，意味着未来平台的点击量会提高。所以pageview和reg_input_success相当于速度与加速度的关系。

3.6浏览量新特征

长时间段规律特征	短时间段规律波动特征
2个月内一星期之前的流量量总和	最近两星期的流量均值
2个月内一星期之前的流量量均值	最近两星期的流量方差
2个月内一星期之前的流量量方差	最近两星期同一类别日的流量最大值
2个月内10天之前的流量量总和	最近两星期同一类别日的流量最小值
2个月内10天之前的流量量均值	最近两星期同一类别日的流量均值
2个月内10天之前的流量量方差	当天要预测是一周中哪一天
2个月内两星期之前的流量量总和	最近两星期的注册成功的人数总和
2个月内两星期之前的流量量均值	最近两星期的注册成功的人数方差
2个月内两星期之前的流量量方差	最近两星期的注册成功的人数均值

3.7 细分群体



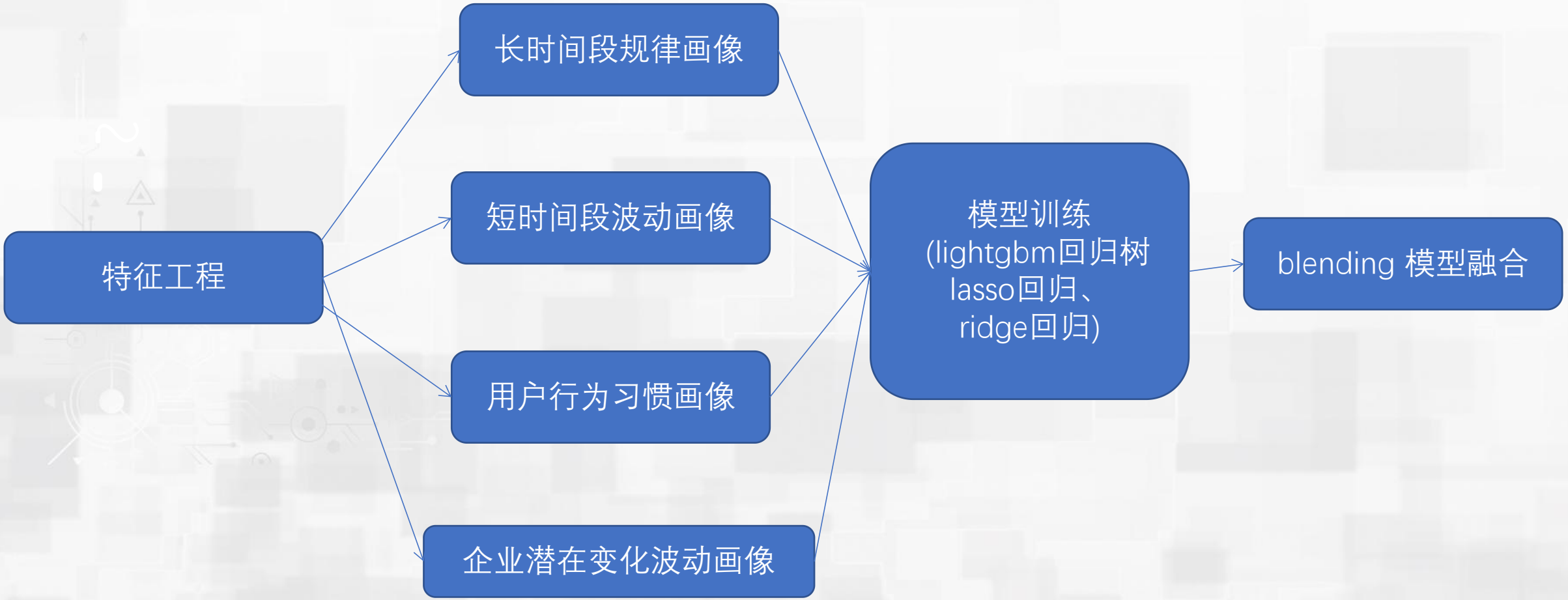
3.8注册量变化分析

	$title	title_all_count	title_count_week1	title_count_week2
0	易观方舟_大数据用户行为分析平台_助力企业精益成长	140213	72197.0	68016.0
1	易观方舟_注册	61568	31887.0	29681.0
2	易观方舟_产品_标签增补_构建用户全景画像，让决策更有依据	17968	4643.0	13325.0
3	看板详情_看板_易观方舟	17451	9549.0	7902.0
4	易观方舟_Argo_易观方舟社区版_方舟CE	15389	7722.0	7667.0
5	易观方舟_产品_分析_多模型灵活自定义实时查询，任意节点数据交互式下钻	11845	5062.0	6783.0
6	易观方舟	10446	6487.0	3959.0
7	易观方舟_行业Demo_了解易观方舟在具体业务场景中的价值	9601	4819.0	4782.0
8	易观方舟_产品_分群_分析模型下钻保存分群，行为和属性自定义分群	6372	2873.0	3499.0
9	易观方舟_Argo下载_易观方舟下载_方舟CE下载	5905	2906.0	2999.0
10	数据分析里的细分维度 - 博客_易观方舟	5258	4132.0	1126.0
11	登录_易观方舟	4941	2877.0	2064.0
12	博客_易观方舟 – 洞见数据背后的故事	4612	1484.0	3128.0
13	分群_易观方舟	4558	2414.0	2144.0
14	易观方舟_公司介绍	4557	2147.0	2410.0
15	电商用户生命周期价值及运营策略 - 博客_易观方舟	4554	2463.0	2091.0
16	易观方舟_电商行业解决方案	4506	2194.0	2312.0

结论：企业最近两周旧用户服务压力增大，

难以更好的引导新(未来)企业用户加入

3.9整体模型展示


特征工程
长时间段规律画像
短时间段波动画像
用户行为习惯画像
企业潜在变化波动画像
模型训练
(lightgbm回归树
lasso回归、
ridge回归)
blending 模型融合

3.9 lightgbm回归树模型

利用lightbm回归树模型进行回归预测

评估指标如下：

$$\sigma_i = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\left(FC_{i,t} - T_{i,t}\right) / \left(T_{i,t}\right)\right)^2}$$

$$F = \sum_{i \in Target} \sigma_i$$

3.9 Ridge、Lasso回归模型

Ridge回归是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，回归用的是L2正则化，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法。即更有效得处理波动数据

Lasso回归用的是L1正则化,是一种压缩估计。它是个精炼的模型，压缩一些回归系数，同时设定一些回归系数为零。因此保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计。

04 实用性总结

- 模型创新性与实用性总结

模型创新、实用性总结

模型创新、实用性总结：

1.模型基于长时间段历史规律与最近短时间段波动变化规律，兼顾全局变化与局部变化符合真实变化

2.模型借用blending,从吸收各个模型的优点，从不同的角度去挖掘变化的规律

3.模型使用基于预测时间段的验证方法,过滤节假日等带来异常数据噪声

4.模型基于大量真实数据变化方向，细分群体进行分析，过滤异常数据群体，兼顾全局变化与局部变化，有利于后期定位对应群体，更快速的调整战略方向，避免不必要的损失，及时抓住相应爆发机会

感 谢 观 看

Business Summary Plan project training template

队伍：橘冯jin