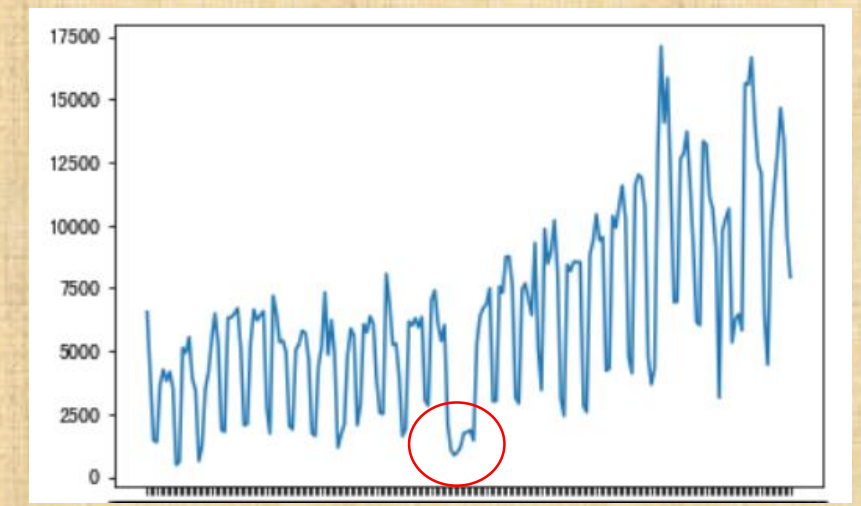
易观×CSDN算法大赛

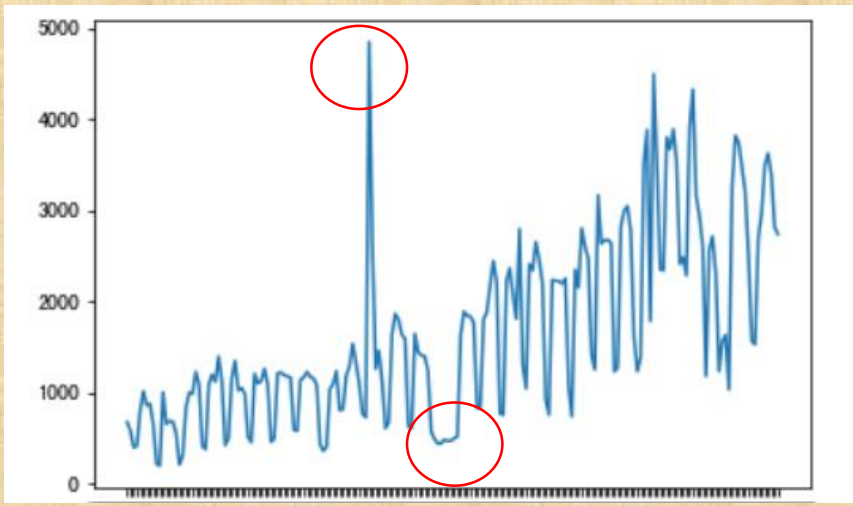
pv、uv流量预测

XYZ　2019/09/29

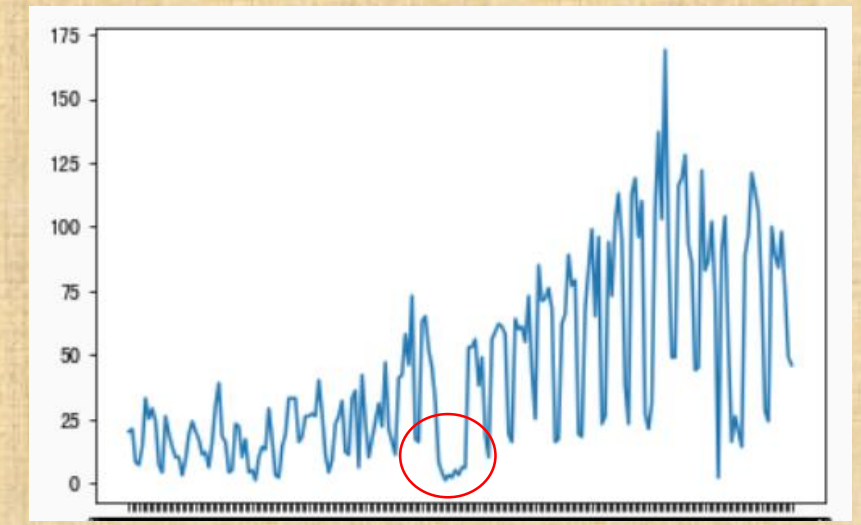
目录



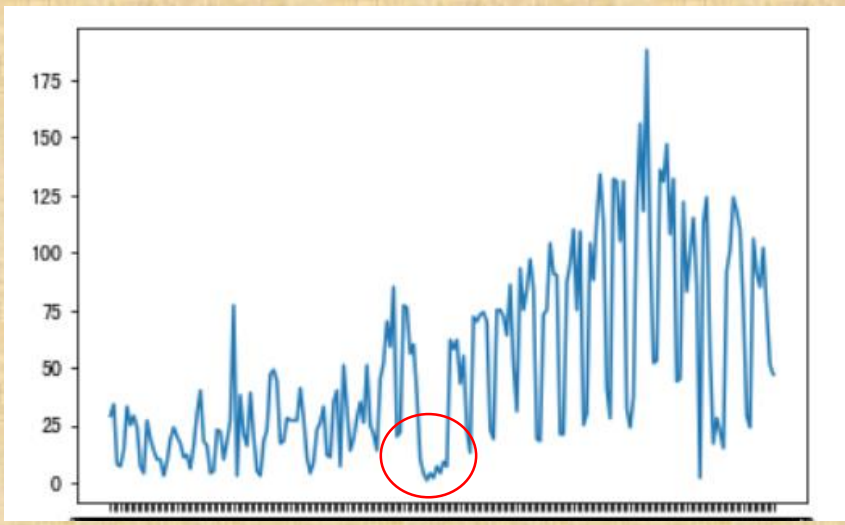
$pageview pv



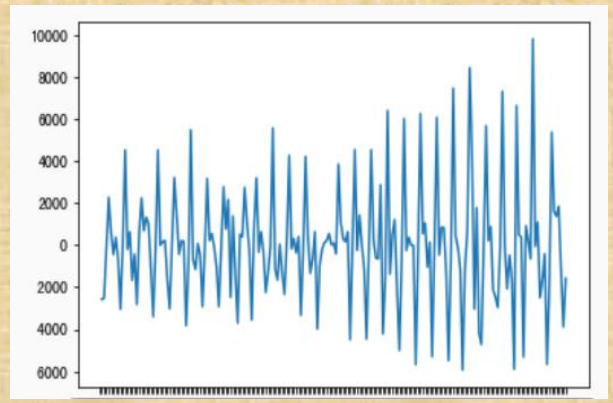
$pageview uv



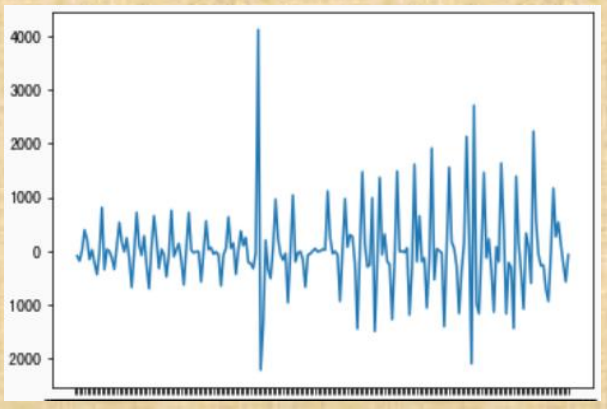
$reg_input_success pv



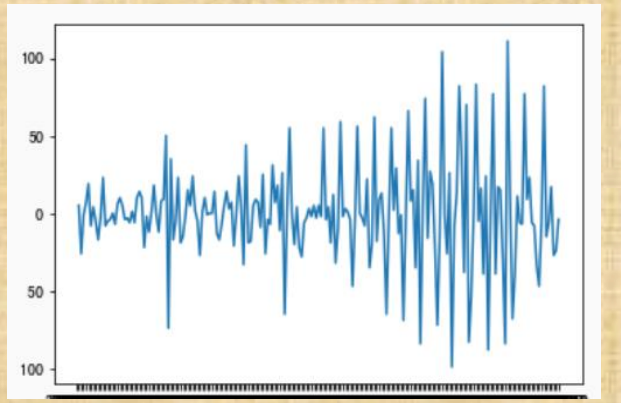
$reg_input_success uv



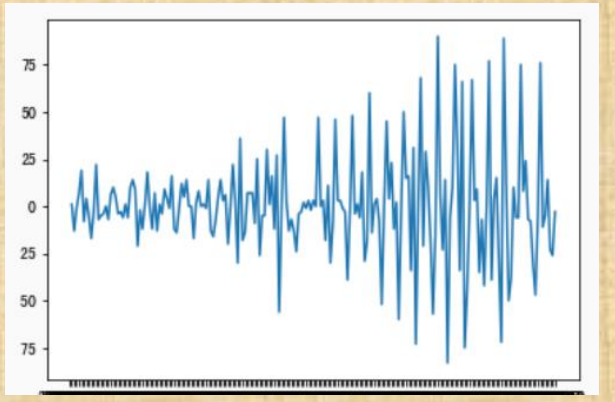
$pageview pv_diff



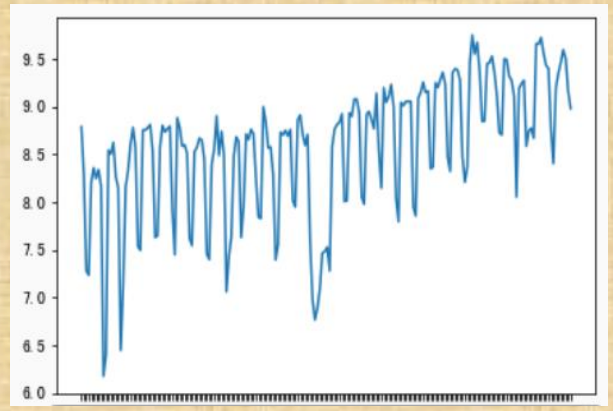
$pageview uv_diff



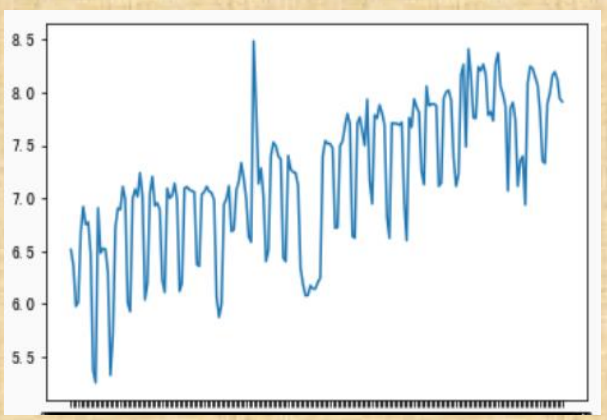
reg_input_success pv_diff



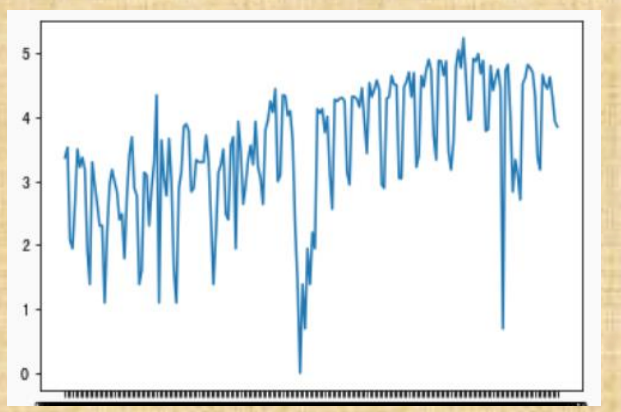
reg_input_success uv_diff



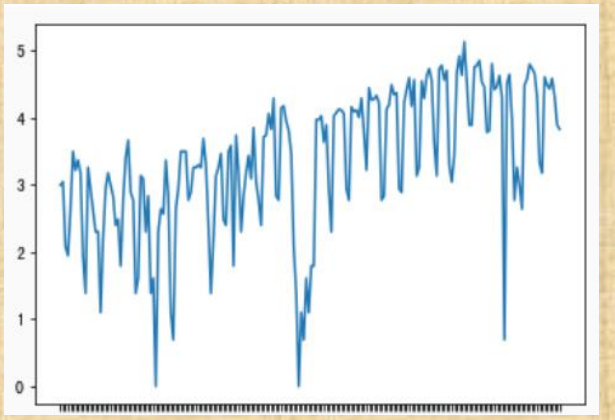
$pageview pv_log



$pageview uv_log



reg_input_success pv_log



reg_input_success uv_log

数据特点

- 周期性
- 上升趋势
- 工作日数据值大于节假日
- 同事件类型下pv、uv走势相似
- 都有明显异常数据数据段（春节）

数据清洗

- 一般异常点处理

$$x_t = \frac{1}{2}(x_{t-7} + x_{t+7})$$

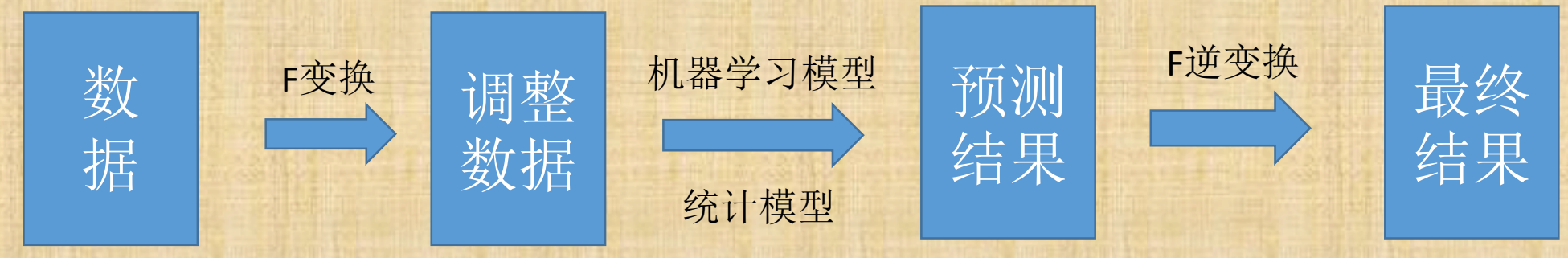
- 一般节假日处理（春节以外）

  按上述公式将数据调整成 '5+2' 的形式

- 春节数据处理

  删去2019年2月4日至2019年2月10日的数据

模型建立



F变换:

$$x_t \rightarrow F(x_t, x_{t-1}, \ldots, x_{t-n}, t)$$

F逆变换:

$$x\_pred_{t+1} \rightarrow F^{-1}(x_t, x_{t-1}, \ldots, x_{t-n}, t+1)$$

假定预测结果单位时间增长率为α，做如下近似:

$$F^{-1}(x_t, x_{t-1}, \ldots, x_{t-n}, t+1) \sim (1+\alpha) F^{-1}(x_t, x_{t-1}, \ldots, x_{t-n}, t)$$

模型建立

可选F变换：

- 差分变换

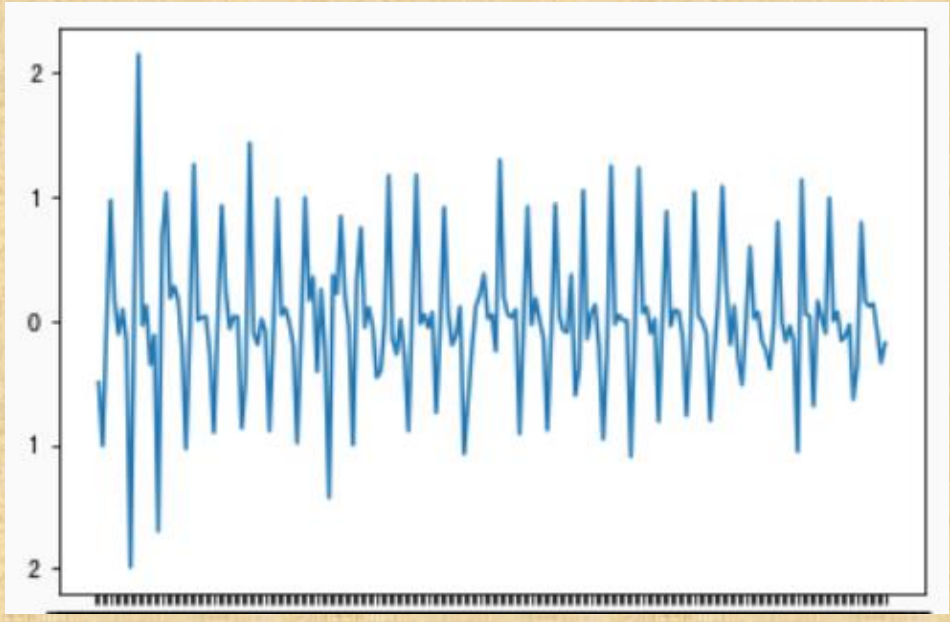
$$F(x_t) = x_t - x_{t-1}$$

- 对数变换

$$F(x_t, t) = \ln x_t - \varepsilon t$$

- 对数-差分变换

$$F(x_t, x_{t-1}, t) = \ln x_t - \varepsilon t - (\ln x_{t-1} - \varepsilon(t-1)) = \ln\left(\frac{x_t}{x_{t-1}}\right) - \varepsilon$$



$pageview pv_log_diff

模型建立

按事件类型和预测流量可以分为4大类：

$pageview-pv、 $pageview-uv、reg_input_success-pv、reg_input_success-uv

一共有28个待预测数据，所以该问题可以看出时4个大问题和28个小问题，理论上应该用28个模型预测

由于LSTM在预测后几天的时候累计误差比较大，实际效果不太理想，所以我们采用按时间划分，前3天用LSTM预测，后4天用ARIMA预测

Simple LSTM

Input:

['pv','uv','reg_submit_click','login','$startup','reg_code_input','$is_first_day','$is_login']

(以上事件相关参数为event_detail.xwhat按天累和,'$is_first_day', '$is_login'为求均值)

#时间长度

n_steps = 20

#每个隐藏层节点数

hidden_size = 128

#LSTM layer层数

layer_num = 2

ARIMA

- 做差分获得平稳时间序列
- 利用statsmodels计算最佳阶数

order = st.arma_order_select_ic(train,max_ar=3,max_ma=3,ic=['aic','bic','hqic'])

得到order.bic_min_order

- 预测数据
- 还原数据

注：这里的数据采用的是每周相同天的数据即 $X=\{x \mid x_{t-7k}, k \in N\}$

偏差分析

数据有明显上涨的趋势，结合平台访问情况课题背景，数据包含两部分：

1. 老用户
2. 新增用户

其中，新增用户和老用户行为操作没有直接关系，因此用历史数据很难预测。同时，用历史数据训练时，也会受到当时新增用户数据的干扰。

谢谢观看