所在组别	2021 年中国高校大数据挑战赛	参赛编号
研究生组		bdc210973

基于 prophet 算法和 logistic 回归的智能运维异常检测与趋势预测

摘要

信息时代化的到来，智能运维的检测和预测变得更加具有现实意义。指标异常检测是智能运维领域的一大热点，其目的在于通过和服务器直接相关的 KPI（关键效益指标）的分析，在时间序列上的异常点来对服务器的运行状态和未来趋势进行分析和构建模型，然后通过告警告知运维人员相关风险，为相关人员提供指导和参考。

针对问题一：针对三项核心指标，存在明显的每日和每周的周期性变化，并不存在长期趋势，因此利用 Prophet 算法求解最优预测和最坏预测，取 0.95 置信区间，在置信之外的点将其归纳为异常值点。将异常个数大于 3 个的区间认定为异常区间。最终识别出整体的孤立点大概在 40 个左右，不同的 KPI 周期聚集在 3-6 个，均以天作为最小异常周期。

针对问题二：利用问题一找到的异常值作为目标变量，异常预测的因变量为哑变量，将问题转化为二分类问题进行求解，Logistic 回归模型是根据离散型或者连续型自变量来分析、预测离散型自变量的回归分析方法，因此基于 logistic 回归建立模型。首先将 67 项 KPI 作为权重指标，对这 67 个自变量进行逐步回归消除变量间的多重共线性，共删除了 11 个变量分别是下行 PhysicalResourceBlock 被使用的平均个数、小区内的最大用户数、ERAB 建立尝试总次数、eNodeB 间同频切换出成功次数、eNodeB 内异频切换出成功次数、使缓存为空的最后一个 TTI 所传的下行 PDCP 吞吐量比、扣除使下行缓存为空的最后一个 TTI 之后的数传时、最大激活用户数、用户随机接入时 TA 值在区间 1 范围的接入次数、用户随机接入时 TA 值在区间 2 范围的接入次数、小区 QCI 为 1 的 DRB 业务 PDCPSDU 下行空口发送的总包数。将剩下的变量构建基于 L1 正则 Logistic 回归模型，最终模型的 F1 值达到 0.7652。

针对问题三：利用问题一构建的时间序列模型，针对每一个小区重新构建 prophet 序列模型。选择模型颗粒均为小时，模型周期为天作为预测周期进行趋势预测，进行回归曲线的拟合，最终获得模型准确度评价 MAPE 取值大约在 9.5 左右。

关键词：智能运维；异常检测；Prophet 算法；logistic 回归；趋势预测；

1 引言

1.1 研究背景

网络信息化的飞速发展带动了整个世界经济的发展，促进了人类社会的进步。然而伴随着网络的使用规模越来越大，企业的的 IT 系统变得越来越复杂，传统的人工运维方式已经难以满足企业的智能化管理要求[1]。近年来，随着机器学习、深度学习等技术的发展，智能运维（artificial intelligence for IT operations，AIOps）的概念被提出，将 AI 与运维结合，利用人工智能技术自动监控和管理 IT 业务，提升运维效率。智能运维包含很多关键场景和技术，涉及大型分布式系统的监控、分析、决策等。裴等人[2]将智能运维划分为3个阶段：针对历史事件，包含瓶颈分析、热点分析、故障传播图构建等技术；针对当前事件，包含异常检测、异常定位、故障根因分析等技术；针对未来事件，包含故障预测、趋势预测等技术。其中 KPI 异常检测是互联网服务智能运维的一个底层核心技术，上述大多数运维关键技术都依赖于 KPI 异常检测的结果。

本文通过和服务器直接相关的运营商基站的 67 个 KPI 性能指标进行分析，来对服务器的运行状态和未来趋势进行预测。其中，数据是从 2021 年 8 月 28 日 0 时至 9 月 25 日 23 时共 29 天 5 个基站覆盖的 58 个小区对应的 67 个 KPI 指标。本研究主要选取三个核心指标进行分析。三个核心指标如下：第一个指标：小区内的平均用户数，表示某基站覆盖的小区一定时间内通过手机在线的平均用户人数；第二个指标：小区 PDCP 流量，通过小区 PDCP 层所发送的下行数据的总吞吐量(比特)与小区 PDCP 层所接收到的上行数据的总吞吐量(比特)两个指标求和得到，表示某基站覆盖的小区在一定时间内的上下行流量总和；第三个指标：平均激活用户数，表示某基站覆盖的小区在一定时间内曾经注册过无线网络的平均人数。通过以这三个核心指标为中心的异常值的检测和趋势预测，给相关人员可以提供具有一定指导和参考意义的方法。

1.2 研究方法

本文通过和服务器直接相关的运营商基站的 67 个 KPI 性能指标进行分析，对服务器的运行状态异常检测和对未来趋势进行预测。涉及到主要的算法 Prophet 算法和 logistic 回归。简单介绍一下这两种方法：

（1）2017 年 Facebook 发布了 Prophet 时序先知模型，主要研究时序数据的特征和时序变化规律，同时预测未来的走势。Prophet 的核心是分析包含周期性、趋势性以及节假日效应等的时间序列特征，它是一种针对时序的有效集成解决方案。Prophet 框架主要是由 4 个组件而构成 1 个加法模型，如下式所示：$y(t) = g(t) + s(t) + h(t) + \epsilon$其中：$y(t)$为时间序列在时间 t 的观测值；$g(t)$为增长函数，它模拟目标序列的一种变化趋势；$s(t)$为以加法形式实现灵活组合各种季节性变化趋势，其还可以通过对数变换适应乘法季节性；$h(t)$为一个比较特殊的组件，有效纳入了不规则假期或特殊事件对预测值的影响贡献值，使得将可预计发生的特殊影响事件作为先验知识融合；ϵ为假设其服从正态分布的噪声因子。Prophet 仅使用时间作为自变量，将时间作为组件的几个线性和非线性函数，明确解释了目标序列的时间依赖结构。Prophet 算法应用广泛，比如梁等人[4]用 Prophet 人工智能算法研究与预测移动通信网络“潮汐效应”现象，其结果具有一定的经济效益；还有张等人[5]Prophet 混合模型应用于基站网络流量长期预测，使得模型具有一定的准确度和鲁棒性。

（2）Logistic 回归模型是根据离散型或者连续型自变量来分析、预测离散型自变量的回归分析方法，是常见的用来处理定性变量的统计方法之一，目前在生物、医学、经济等领域应用十分广泛。例如 1838 年 Verhuist[6]首次提出 logistic 概率函数又称增长函数，并且他用 logistic 作增长曲线，利用该曲线进行人口统计学的研究。之后，Logistic 回归完美解决了线性回归无法解决事件因变量为连续变量时的假设条件问题，

加之计算机的普及与发展，使得 Logistic 回归在之后发扬光大，广泛应用于各行各业的预测，尤其是经济学领域和医学领域。Logistic 回归分为三大类[7]，一种是因变量为二分类 Logistic 回归，这种回归叫做二因素 Logistic 回归，因变量结局只有是或者不是，比如本研究中的预测三个指标“异常”或“非异常”。第二种是因变量为无序多分类的 Logistic 回归，这种回归方式用于预测较少，因为结局的无序及不明确性，不适合用于预测，这种回归叫做多因素 Logistic 回归。还有一种是因变量为有序多分类的 Logistic 回归，比如高血压的程度是高、中、低等可分等级的结局，这种回归也叫累计 Logistic 回归。很明显，由于本题是预测三个指标是否异常两种情况，即事件发生仅能用 1（异常）或 0（非异常）来评判，并未有程度分级，所以本研究所用为二分类 Logistic 回归。

1.3 模型假设

根据题意，我们做出如下假设：

（1）本题中所有小区之间仅存在物理性能上的 KPI 差异，不存在人为的外来影响因素和节假日活动、政策的影响。

（2）本题中服务器本身的运作是稳定的，不存在异常的流量访问导致的异常。

（3）本题中 58 个小区之间的区别仅仅在于 KPI 的性能和小区人员的日常使用习惯和人数问题，不存在其他的影响因素

1.4 研究意义

异常检测（异常诊断/发现）、异常预测、趋势预测，是智能运维中首当其冲需要解决的问题。针对传统模型无法对网络流量异常进行准确识别和检测的问题，提出一种基于时间序列分析的网络流量异常检测模型．首先提取网络流量的原始数据，并对原始数据进行去噪处理，消除干扰因素的影响；然后采用时间序列分析法挖掘网络流量数据之间的变化关系，建立网络流量异常检测模型。本题通过分析小区平均用户数、小区 PDCP 流量、平均激活用户数的时序数据的特征与趋势，进一步分析小区平均用户数、小区 PDCP 流量、平均激活用户数的预测准确度，并通过 Prophet 算法、Logistic 回归等方法对小区流量异常情况进行预测。基于时间序列预测模型可以准确、及时地检测网络流量的异常行为，我们不仅可以判断先有数据的异常点和异常周期，也可以对未来一段时间的异常点和异常周期进行判断并预测，这对于小区及时调整流量资源配置具有重要的参考意义。

2 问题一分析

2.1 问题分析

首先，根据三个指标我们分别按照月、周、天为时间周期对数据做了一些整体性的描述，结果如下图 1-3。我们发现：

（1）数据按月拆分来看，小区内的平均用户数从月初到月末有整体轻微下降的趋势，下降幅度大概在 6%左右，但并未存在周期趋势。

（2）数据按周拆分来看，很明显存在周期性的变化。

（3）数据按天拆分来看，也可以看出存在周期性的变化。

还对小区内的平均用户数、小区 PDCP 流量、平均激活用户数三个关键指标在时间轴上的变化分别进行了分析如图 4，可以看到明显的周期性变化。

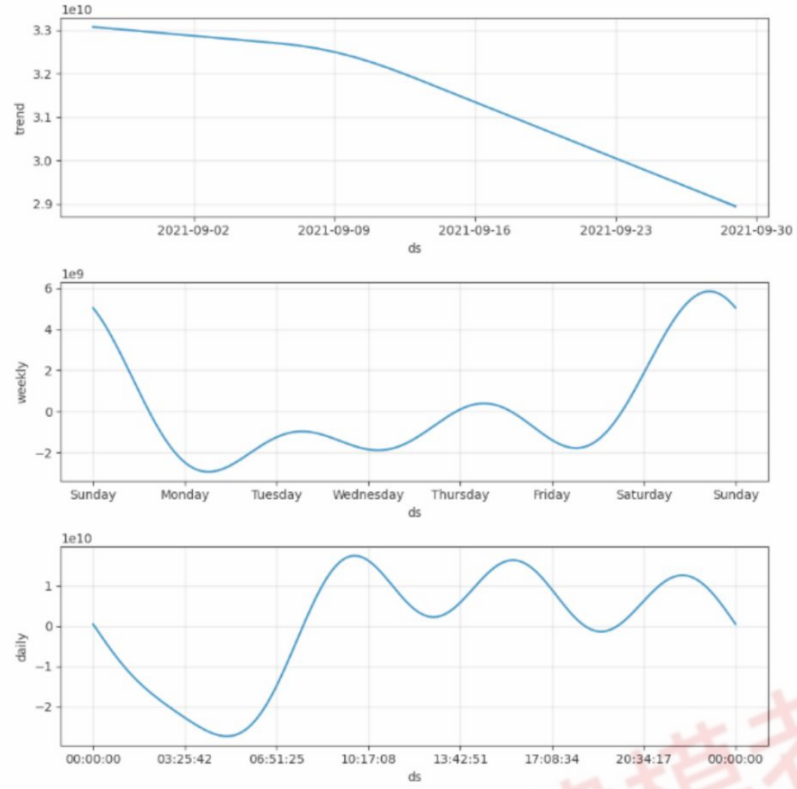

图 1. 小区内的平均用户数月、周、天周期描述

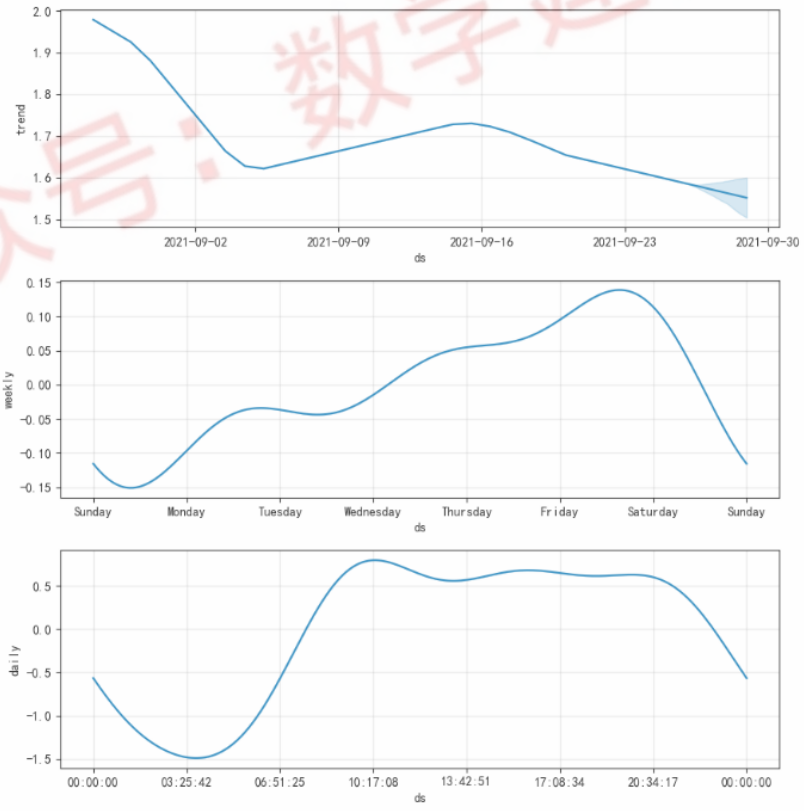

图 2. 平均激活用户数月、周、天周期描述

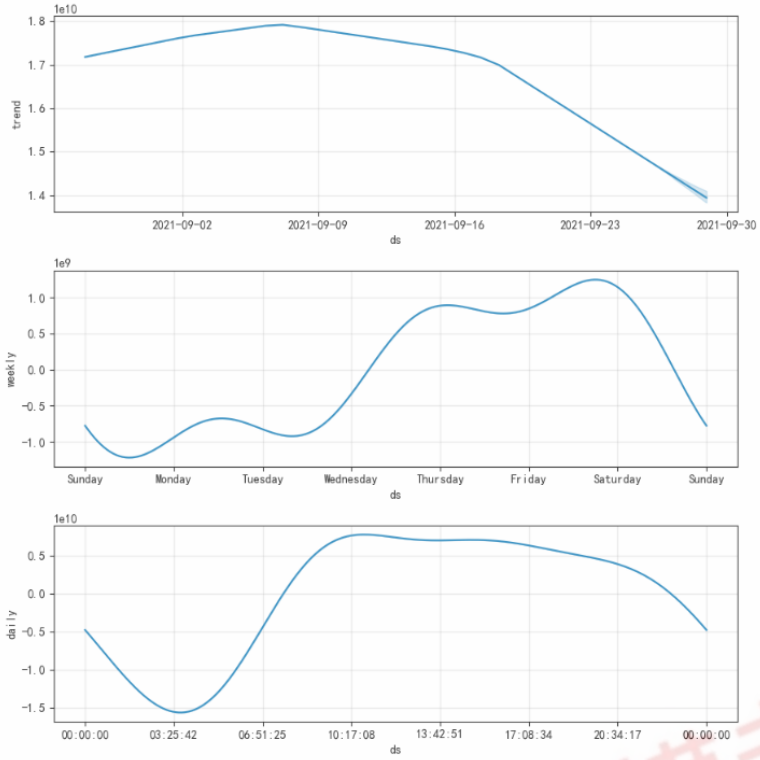

图 3. 小区 PDCP 流量月、周、天周期描述

图 4.三个指标周期性描述

首先使用 spearman 相关系数得到三个指标间的热力图，从图 7 可以看出三个关键指标是强相关性的，因此不能分开单独做异常检测。其次通过可视化的方式查看小区 PDCP 流量时间序列是否存在长期的趋势性规律或者变化，可以看出并不存在月级别的变化。抽取其中一个小区的小区 PDCP 流量绘制下图 5，对比将 58 个小区整合在一起的小区 PDCP 流量图 6，能够很明显的看出每个小区之间存在着不同的小时变化趋势，但是小区整体而言，更加明显的趋势是每天的变化趋势作为整体变化。

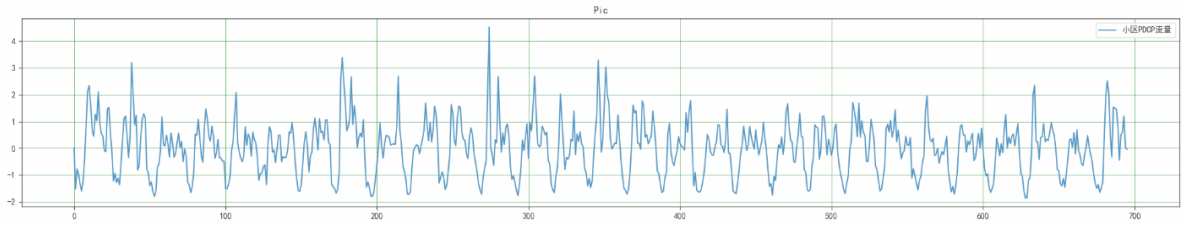

图 5.其中一个小区 PDCP 流量

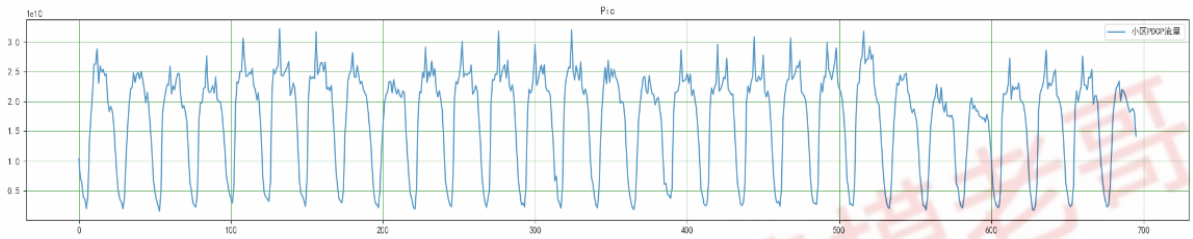

图 6.小区整合在一起 PDCP 流量

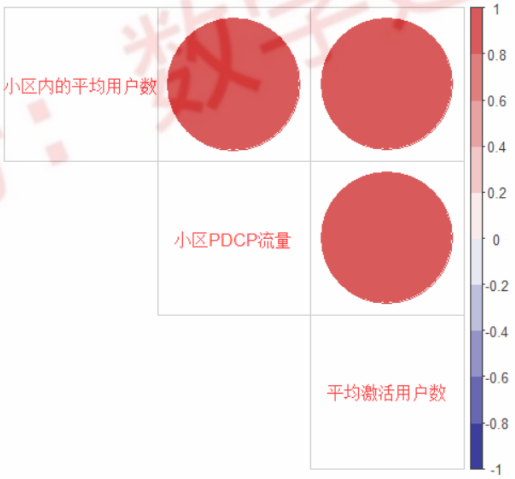

图 7.三个指标之间的相关性热力图

2.2 模型建立

根据上述的描述性图表分析可以看出三个核心指标存在明显的每日和每周的周期性变化，并不存在长期趋势。因此利用时间序列分解算法将时间序列进行分解，运用到的模型是 Prophet 模型，此模型的原理是：

$$y(t) = g(t) + s(t) + h(t) + \epsilon \quad (1)$$

其中$g(t)$是趋势函数，$s(t)$表示周期性函数，$h(t)$表是节假日、假期函数，ϵ表示误差或者是噪声等。

（1）$g(t)$利用的是基于逻辑回归函数的非线性增长。类似于服务器的 KPI 增长指标的变化趋势存在大量的约束条件，因此很难达到线性增长的能力，由自变量的变化

可知，自变量的变化幅度明显要比因变量要大，因此变化的增长幅度更倾向于非线性增长。

$$g(t)=\frac{C(t)}{1+exp(-(k+a(t)^T\delta))\cdot(t-(m+a(t)^T\gamma))} \quad (2)$$

$$a(t)=(a_1(t),\cdots,a_S(t))^T,\delta=(\delta_1,\cdots,\delta_S)^T,\gamma=(\gamma_1,\cdots,\gamma_S)^T \quad (3)$$

其中，$C(t)$表示的是一个承载量，限定了函数随时间变化的最大上下限。a(t) 一共有 s 个变量就是就是 s 个时间戳，对应的（第二个变量）,就是对应时间戳上面的增长率的变化量，最终利用偏移量 m 能够用来调整两个时间戳之间的分段。

（2）$s(t)$是变量的增加是由于时间序列当中可能包含多种天、周、月等周期类型的季节性趋势，很明显能够看出本题当中的三个最重要 KPI 指标当中在存在高度相关性的情况下仍然存在 24 小时之内的峰谷变化以及在一个星期当中周末的流量明显要比工作日要高的规律性情况，因此有必要引入季节性趋势。使用傅立叶级数来模拟时间序列的周期性：假设 P 表示时间序列的周期 P = 7 表示以周为周期。傅立叶级数的形式：

$$s(t)=\sum_{n=1}^{N}\left(a_n cos\left(\frac{2\pi nt}{P}\right)+b_n sin\left(\frac{2\pi nt}{P}\right)\right) \quad (4)$$

其中 N 表示希望在模型中使用的这种周期的个数，较大的 N 值可以拟合出更复杂的季节性函数，然而也会带来更多的过拟合问题。

（3）$h(t)$节假日效应是为了防止节假日这种单独的重大时间对时间序列的影响，但是这种影响往往又不具有周期性，因此如果出现这种情况增加节假日项是非常有必要的。为了表示节假日效应，需要考虑节假日的影响程度和影响的时间长度不同，因此需要设定不同的时间窗口D_i，对于出现的第 i 个节假日D_i代表该节假日前后的一段时间。另外需要一个相应的指示函数，同时需要一个参数k_i来表示节假日产生的影响范围。假设有 L 个节假日，那么节假日效应模型就是：

$$h(t)=Z(t)k=\sum_{i=1}^{L}k_i\cdot 1_{t\epsilon D_i} \quad (5)$$

其中$Z(t)=(1_{t\epsilon D_i},\dots,1_{t\epsilon D_L})$和$k=(k_1,\dots,k_L)$。

2.3 模型求解

附件一所提供的数据为(58*29*24,71)的数据维度表，能够很明显的分辨出 58 个小区在 24*29 的时间段内每个时间都具有相对应的 KPI 维度指标，而且不存在数据缺失情况。因此针对第一问，通过将 58 个小区进行合并求取一个平均值作为(24*29,71)组序列进行回归函数的构建，其中核心函数参数设定见表 1。

表 1.设定核心函数参数

Prophet()函数主要参数	
Changepoint_prior_scale（设置拟合跟随性）	0.05
Growth	logistic
Daily_seasonality（设置日规律拟合）	True
Weekly_seasonality（设置周规律拟合）	True
Seasonality_mode	Multiplicative
Interval_width（设置置信区间值越小，上下限的带宽越小）	0.95
Make_future_dataframe()函数的主要参数	
Periods（设置预测长度）	24*3
Freq（设置最小时间颗粒）	H

通过可视化判别的方式来探索数据集的规律和关联，得到三个指标异常点及异常周期如下结果图 8-10：

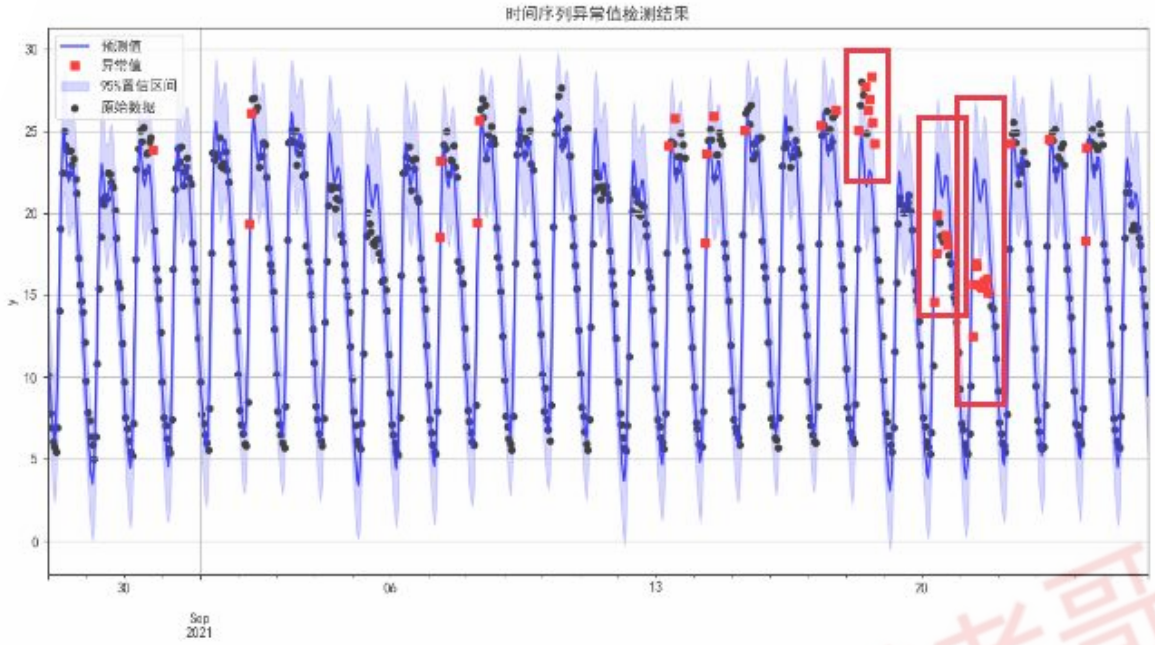

图 8.小区平均用户数异常点及周期识别

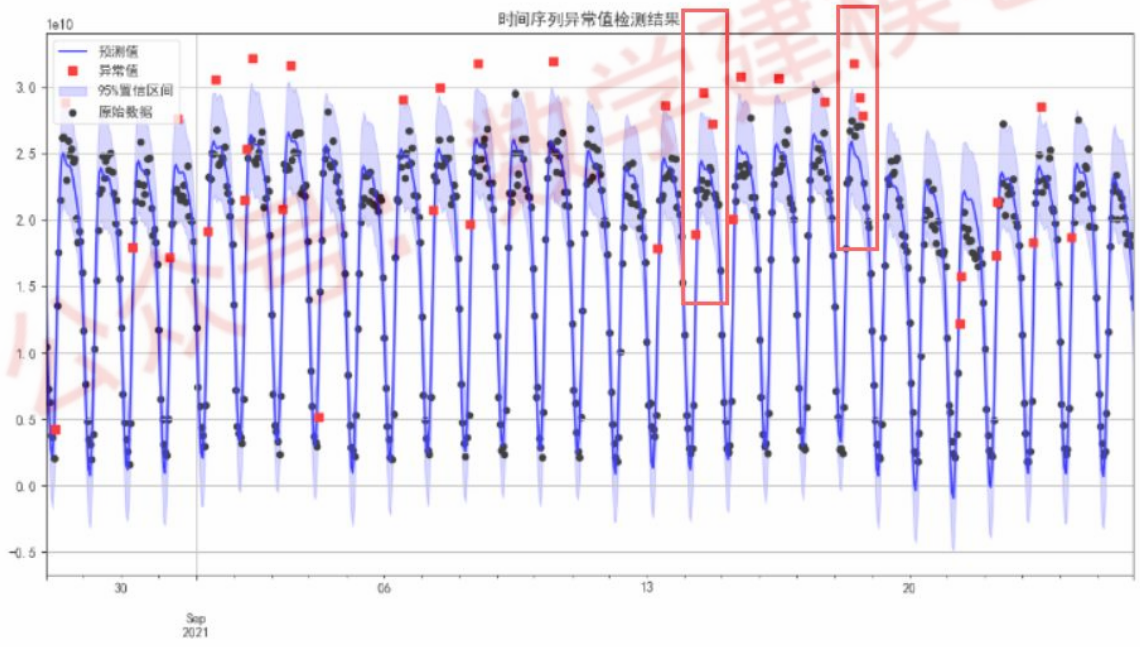

图 9. 小区 PDCP 流量异常点及周期识别

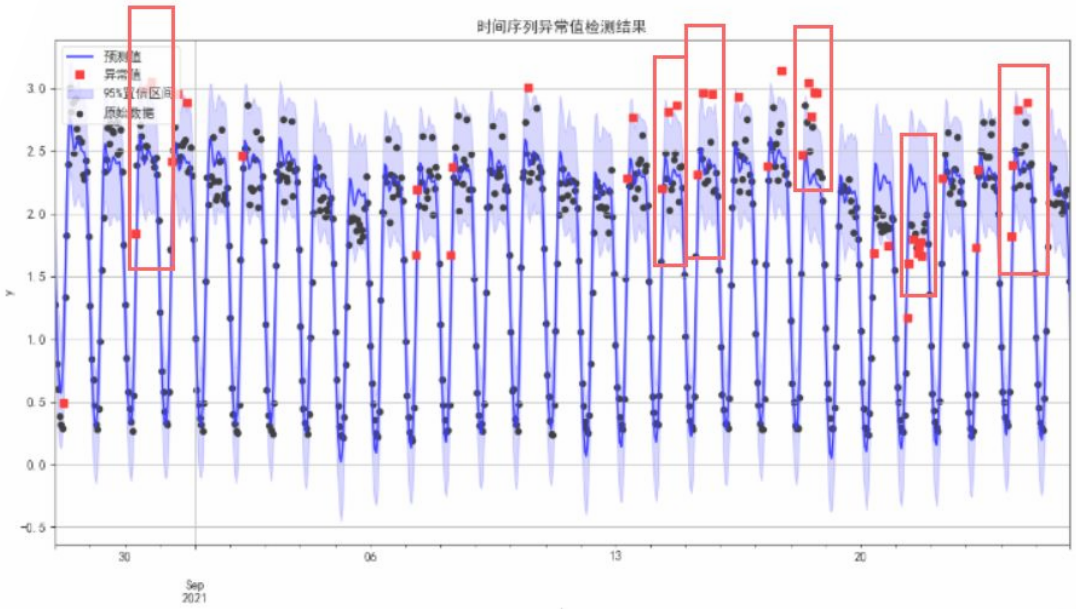

图 10. 平均激活用户数异常点及周期识别

通过限定 95%的置信区间的预测方式，我们获得了最优预测和最坏预测以及最终的预测曲线，并将不在曲线之内的参数点判定为异常值，最终通过限定不同的核心 KPI 获得不同的异常点孤立点个数，另外通过一天作为一个周期，认定一天之内超过三个异常点的情况被认为是一个异常周期，最终获得结果表：

	时间周期选择标准	异常孤立点的个数	异常周期个数
小区内的平均用户数	天	42	3
小区 PDCP 流量	天	38	2
平均激活用户数	天	47	6

3 问题二分析

3.1 问题分析

本题数据来源基于题目一，问题二也是以天为划分单位。为了满足后续研究的需求，首先进行特征的选择，对数据进行多重共线性的消除；其次考虑到异常预测为因变量是 0,1 的二分类问题，因此使用常见的 logistic 回归模型，为防止过拟合在此基础上加入了 L1 正则项来预测是否异常，将因变量有异常记为 TURE=1，无异常记为 FALSE=0。

由于 logistic 回归对多重共线性的数据比较敏感，因此需要消除自变量间的多重共线性。一般来说如果存在不全为零a+1 个数c_0，c_1，……，c_a，使得

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \cdots + c_a x_{ia} \approx 0，\ i=1,2，\cdots，\ n \qquad (6)$$

则说明自变量x_1，x_2，……，x_a之间是存在多重共线性的。诊断多重共线性的方法有很多种，常见计算变量间的共线性有方差扩大因子、特征根和条件指数。本文使

用条件指数判断法，

$$k = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \tag{7}$$

其中λ 是$X^T X$的特征值，R 语言 k 值得到大于 30 的时候，则存在多重共线性。说明数据间存在多重共线性。为了更加直观的看到变量之间的相关性，由于数据集中的变量有定性变量也有定量变量，因此我们使用 spearman 相关系数做出各个变量之间相关图如图 11（由于变量过多，这里只展示前 10 个变量间的热力图）。计算变量间的相关性系数：

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{8}$$

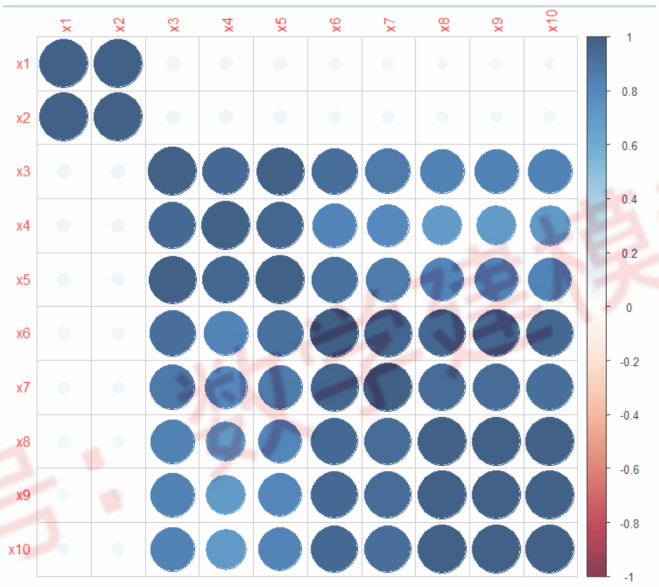

图 11 变量间部分热力图

对 67 个指标（见表 2）进一步使用 SPSS 逐步回归法来消除自变量间的多重共线性，共 11 步删除了 11 个变量见表 3，分别是 X5、X8、X12、X17、X20、X24、X25、X29、X47、X48、X66。

表 2.模型对应指标

时间	X1	空口上报全带宽 CQI 为 5 的次数	X35
上行可用的 PRB 个数	X2	空口上报全带宽 CQI 为 6 的次数	X36
下行可用的 PRB 个数	X3	空口上报全带宽 CQI 为 7 的次数	X37
上行 PhysicalResourceBlock 被使用的平均个数	X4	空口上报全带宽 CQI 为 8 的次数	X38
下行 PhysicalResourceBlock 被使用的平均个数	X5	空口上报全带宽 CQI 为 9 的次数	X39
上行 PUSCH 的 PhysicalResourceBlock 被使用的平均个数	X6	空口上报全带宽 CQI 为 10 的次数	X40
小区内的平均用户数	X7	空口上报全带宽 CQI 为 11 的次数	X41
小区内的最大用户数	X8	空口上报全带宽 CQI 为 12 的次数	X42

RRC 连接建立完成次数	X9	空口上报全带宽 CQI 为 13 的次数	X43
RRC 连接请求次数（不包括重发）	X10	空口上报全带宽 CQI 为 14 的次数	X44
ERAB 建立成功总次数	X11	空口上报全带宽 CQI 为 15 的次数	X45
ERAB 建立尝试总次数	X12	用户随机接入时 TA 值在区间 0 范围的接入次数	X46
ERAB 异常释放总次数	X13	用户随机接入时 TA 值在区间 1 范围的接入次数	X47
ERAB 正常释放总次数	X14	用户随机接入时 TA 值在区间 2 范围的接入次数	X48
系统间切换出 ERAB 正常释放总次数	X15	用户随机接入时 TA 值在区间 3 范围的接入次数	X49
eNodeB 内同频切换出成功次数	X16	用户随机接入时 TA 值在区间 4 范围的接入次数	X50
eNodeB 间同频切换出成功次数	X17	用户随机接入时 TA 值在区间 5 范围的接入次数	X51
eNodeB 内同频切换出执行次数	X18	用户随机接入时 TA 值在区间 6 范围的接入次数	X52
eNodeB 间同频切换出执行次数	X19	用户随机接入时 TA 值在区间 7 范围的接入次数	X53
eNodeB 内异频切换出成功次数	X20	用户随机接入时 TA 值在区间 8 范围的接入次数	X54
eNodeB 间异频切换出成功次数	X21	用户随机接入时 TA 值在区间 9 范围的接入次数	X55
eNodeB 内异频切换出执行次数	X22	用户随机接入时 TA 值在区间 10 范围的接入次数	X56
eNodeB 间异频切换出执行次数	X23	用户随机接入时 TA 值在区间 11 范围的接入次数	X57
使缓存为空的最后一个 TTI 所传的下行 PDCP 吞吐量比	X24	MR 测量上报 RSRP 在 Index0 区间的次数	X58
扣除使下行缓存为空的最后一个 TTI 之后的数传时	X25	MR 测量上报 RSRP 在 Index1 区间的次数	X59
使 UE 缓存为空的最后一个 TTI 所传的上行 PDCP 吞吐量	X26	MR 测量上报 RSRP 在 Index2 区间的次数	X60
扣除使 UE 缓存为空的最后一个 TTI 之后的上行数传	X27	MR 测量上报 RSRP 在 Index3 区间的次数	X61
平均激活用户数	X28	MR 测量上报 RSRP 在 Index4 区间的次数	X62
最大激活用户数	X29	小区 QCI 为 1 的 DRB 业务 PDCPSDU 上行丢弃的总包数包	X63
空口上报全带宽 CQI 为 0 的次数	X30	小区 QCI 为 1 的 DRB 业务 PDCPSDU 上行期望收到的总包数	X64
空口上报全带宽 CQI 为 1 的次数	X31	小区 QCI 为 1 的 DRB 业务 PDCPSDU 下行空口丢弃的总包数	X65

空口上报全带宽 CQI 为 2 的次数	X32	小区 QCI 为 1 的 DRB 业务 PDCPSDU 下行空口发送的总包数	X66
空口上报全带宽 CQI 为 3 的次数	X33	小区 PDCP 流量	X67
空口上报全带宽 CQI 为 4 的次数	X34	是否异常	y

表 3.消除多重共线性

步骤	变量
1	X8
2	X8, X66
3	X8, X66, X49
4	X8, X66, X49, X20
5	X8, X66, X49, X20, X25
6	X8, X66, X49, X20, X25, X39
7	X8, X66, X49, X20, X25, X39, X37
8	X8, X66, X49, X20, X25, X39, X37, X47
9	X8, X66, X49, X20, X25, X39, X37, X47, X48
10	X8, X66, X49, X20, X25, X39, X37, X47, X48, X17
11	X8, X66, X49, X20, X25, X39, X37, X47, X48, X17, X5

3.2 模型建立及求解

Logistic 回归处理二分类问题经典的算法之一，且 logistic 回归易于最优化求解。一般的 logistic 回归模型时常容易欠拟合，通过我们不断的优化，就又会出现过拟合现象。针对过拟合问题，通常会考虑两种方法。第一种是减少特征的数量；第二种是加入正则项。在此我们用到正则化惩罚，即惩罚数值较大的权重参数，降低它们对结果的影响。

在上面处理好的数据集的基础上，进行建模。由于数据集有异常和无异常分布不平衡，会导致我们在建模过程中模型结果会出现偏向的情况，因此我们采用下采样方法平衡我们的数据。然后我们随机选择为70%训练集，30%为测试集。对训练集进行五种不同正则化惩罚力度（0.01,0.1,1,10,100）的 logistic 回归建模，其中：自变量$x = (x_1$，x_2，……，$x_i)^T$，因变量 y，$h_\theta(x)$是 sigmoid 函数（$h_\theta(x) = \frac{1}{1+e^{-\theta^T}}$），基本过程如下：

首先根据 logistic 回归模型的损失函数$cost(h_\theta(x), y)$（时常使用对数损失函数见以下公式）：

$$cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)) & if\ y = 1 \\ -log(1 - h_\theta(x)) & if\ y = 0 \end{cases} \quad (9)$$

其次加上 L1 正则项，得到 L1 正则化 logistic 回归模型的损失函数$J(\theta)$(λ正则化惩罚系数)：

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[-y^{(i)}log(h_\theta(x^{(i)})) - (1-y^{(i)})log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{2m}^{n}|\theta_j| \quad (10)$$

最后使用随机梯度下降法，不断更新θ_j，其中θ_0不参与惩罚，得到最优参数θ，使得损失函数$J(\theta)$最小。

$$\theta_0 := \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\,x_0^i \quad (11)$$

$$\theta_j := \theta_j - \alpha\left[\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_j^i + \frac{\lambda}{m}\theta_j\right] \quad (12)$$

我们从准确率、召回率、精确度、F1、AUC 不同的指标来评估我们模型的性能见图 12（蓝色为测试集，红色为训练集）和表 4，五种不同正则化惩罚力度（0.01,0.1,1,10,100）在惩罚系数为 0.1 的时候效果是最好的。

表 4.不同惩罚系数的评价指标对比

惩罚系数	0.01	0.1	1	10	100
F1	0.7459	0.7652	0.7107	0.7385	0.6817
AUC	0.7307	0.7434	0.7004	0.7219	0.6789
准确率	0.7314	0.7457	0.7	0.7229	0.6771
召回率	0.7419	0.7796	0.6935	0.7366	0.6505
精确率	0.75	0.7513	0.7288	0.7405	0.7160

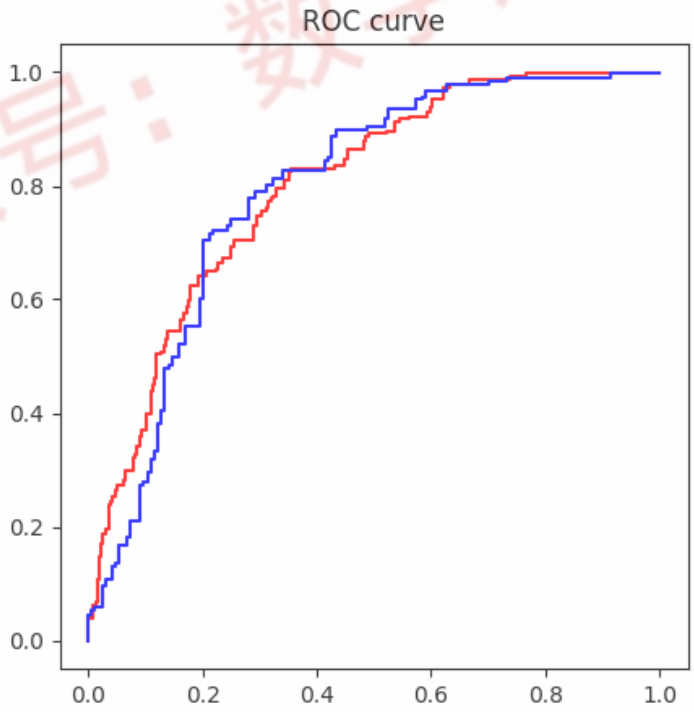

图 12.ROC 曲线对比图

4 问题三分析

4.1 问题分析

针对第一问和第二问给出的说明和分析，我们能够知道每一个小区其实都存在他们自己的一些属性特质和情况，因此最好的做法是依据每一个小区构建一个模型，但是这样的模型就失去了泛化能力。

问题需要我们能够从所给定的条件当中 KPI 对后面三天的情况进行预测，由于没有给定之后的 KPI 的变化条件，因此我们可以通过规律性拟合的方式来给出之后三天的预测状况。

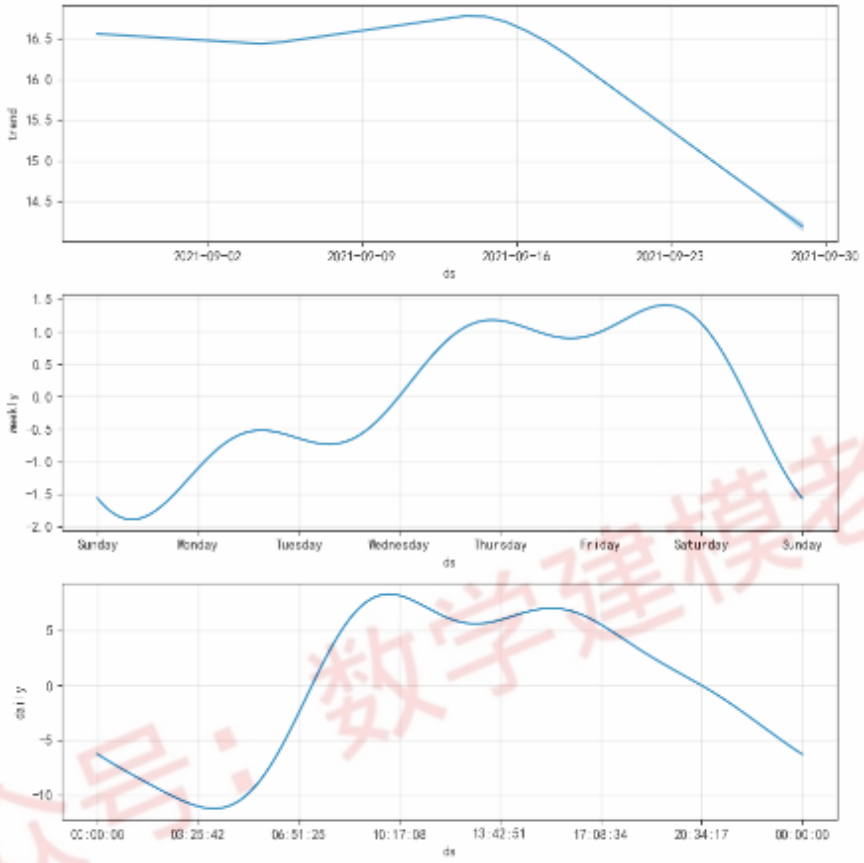

图 13. 小区流量变化趋势

上面图 13 反映了小区内的流量变化趋势，从第二个图和第三个图反映的流量变化可以看出有很明显的周期性，即每一天的流量变化和每周的流量变化有很明显的周期性。第二张图里明显可以看出人们在周四、周五和周六流量使用量明显升高，周一、周日的流量使用量明显低于其余天数，这很符合居民的流量使用现状。

结合日常可以进行如下解释：周一需要开启新一周的工作，所以认真工作的时间远高于使用电子产品的时间，周日是一周的末尾，需要完成这周遗留的工作，所以使用电子产品的时间也不多，导致流量使用下降。第三张图里可以看出在一天中人们在上午十点到下午五点使用流量最多，晚上八点以后使用的流量逐渐减少，直到凌晨三点使用流量降到最低。结合日常也可以进行如下解释：大多数人们在白天使用电子产品的频率高于在晚上使用电子产品的频率。这个结果也很符合现代人们的现状。由图可以看出数据的规律性强，所以对于第三问的求解和预测，仍然使用 Prophet 算法，对规律性强的时间序列的拟合优度更好。对所有的小区整合之后如图 14：

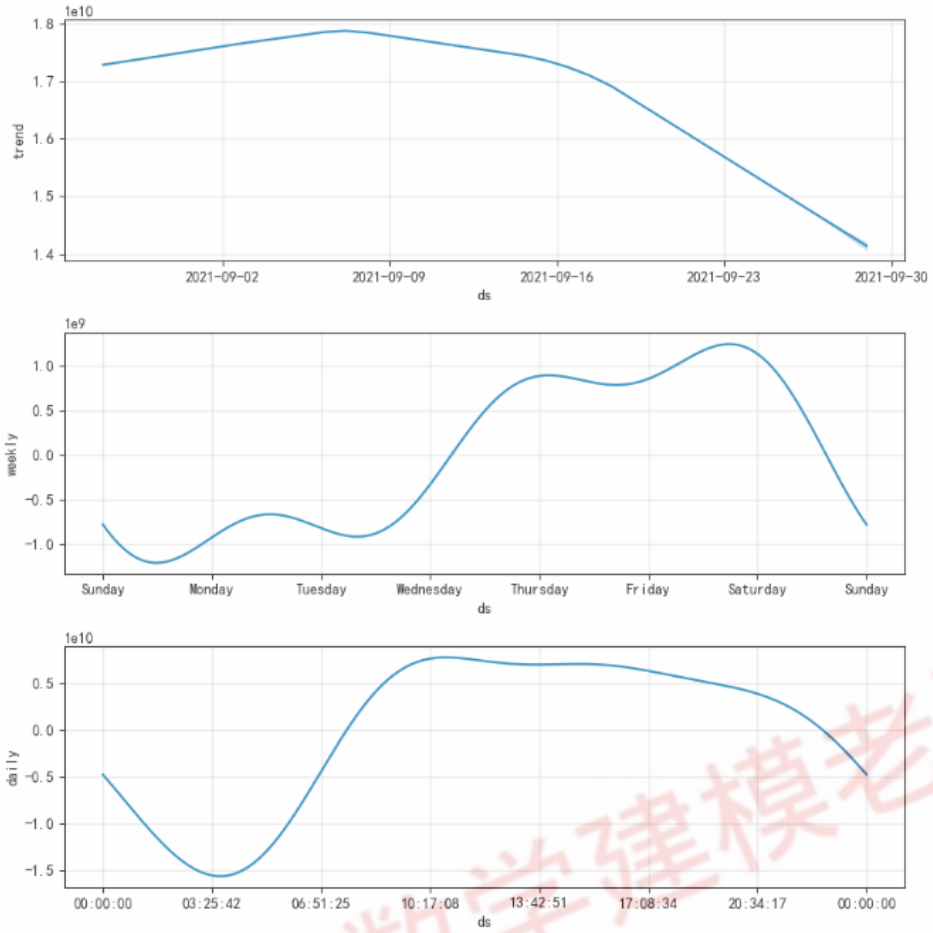

图 14.所有小区整合后趋势图

很明显就会发现对于每天的周期变化变得非常的模糊，说明小区人们每天的生活节奏是不相同的，但是在星期上面仍然存在周期性的特征，说明小区居民每周的流量使用趋势是大致相同的。然后我们就整合所有的小区之后再来进行异常值检测，得到如下图 15 结果：

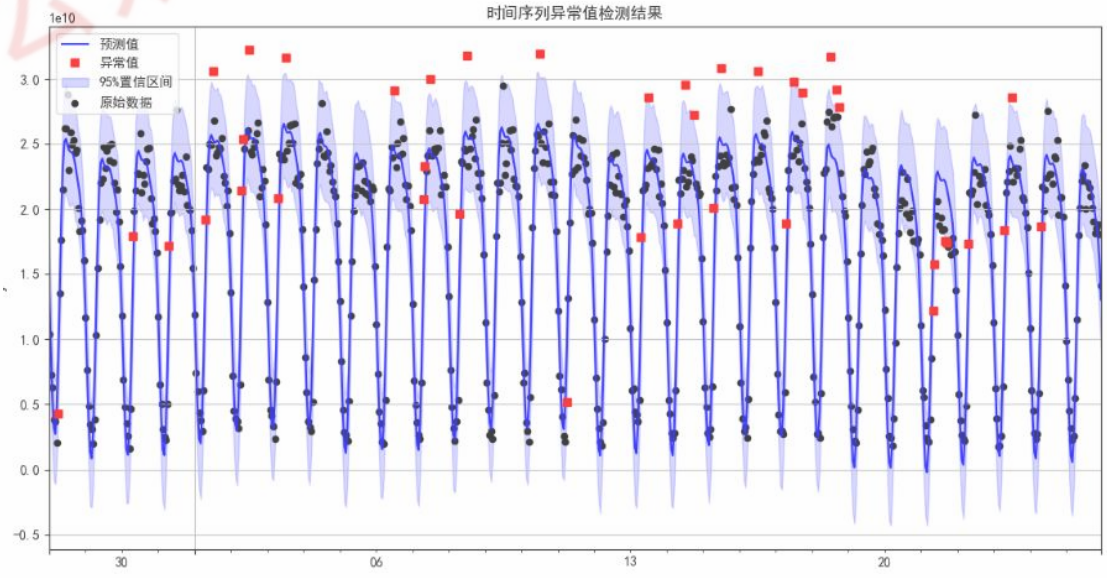

图 15 所有小区异常检测

4.2 问题三模型求解：

对于第三问的求解和预测，仍然使用问题一相同的模型构建方式 Prophet 算法，对规律性强的时间序列的拟合优度更好。为了能够提升模型构建的准确度，针对每一个小区获得更多的信息，对每一个小区进行建模求解。

首先通过构建 Prophet 模型进行对 9 月 26 日-28 日进行预测，其中包括构建待预测日期数据框，在这里设置 periods（代表除历史数据的日期外再往后推）为 24 * 3；其次通过定义 MAPE（平均绝对百分误差）函数，对之前的数据加预测的三天的数据进行模型准确度的分析，最终拟合图如下图 16，最终获得模型的准确度在 9.5 左右。详细预测结果见附录二。

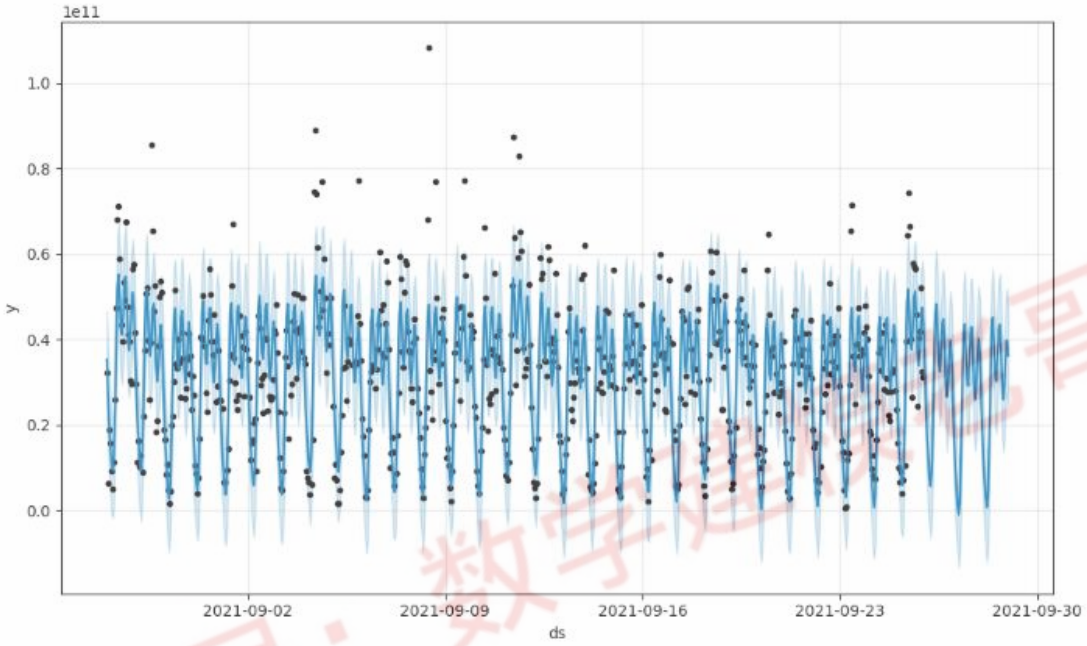

图 16. 拟合图

参考文献

[1] 李青.基于大数据分析的云资源池告警信息关联方案[J].电信科学,2020,36(10):159-171.
[2] 裴丹, 张圣林, 裴昶华. 基于机器学习的智能运维[J]. 中国计算机学会通讯, 2017, 13(12): 67-73.
[3]李丽萍, 段桂华, 王建新. 基于 Prophet 框架的银行网点备付金预测方法 [J].中南大学学报 (自然科学版), 2019,50(1): 75–82.
[4]梁志生,韩永涛,林翔.基于 Prophet 人工智能算法的网络潮汐效应预测研究[J].电信工程技术与标准化,2021,34(09):60-68.
[5]张家晨,左兴权,黄海,韩静,张百胜.Prophet 混合模型应用于基站网络流量长期预测[J/OL].计 算 机 工 程 与 应 用:1-11[2021-11-01].http://kns.cnki.net/kcms/detail/11.2127.TP.20210706.0941.002.html.
[6] P.J. Verhulst. Notice sur lalois Que la Population Suit Dans Sons Acctoissen-ment[J].Corr. Math. Phys. Et Physiyue,1938,(10):113-130.
[7]隋泽森. 通过危险因素对食管癌术后吻合口瘘的预测：Logistic 回归模型与人工神经网络模型的建立及比较[D].南方医科大学,2019.