JDD 人口动态普查与预测比赛

张春光

厦门大学

报告内容

赛题背景

1. 人口普查
2. 通信设备
3. 大数据方法

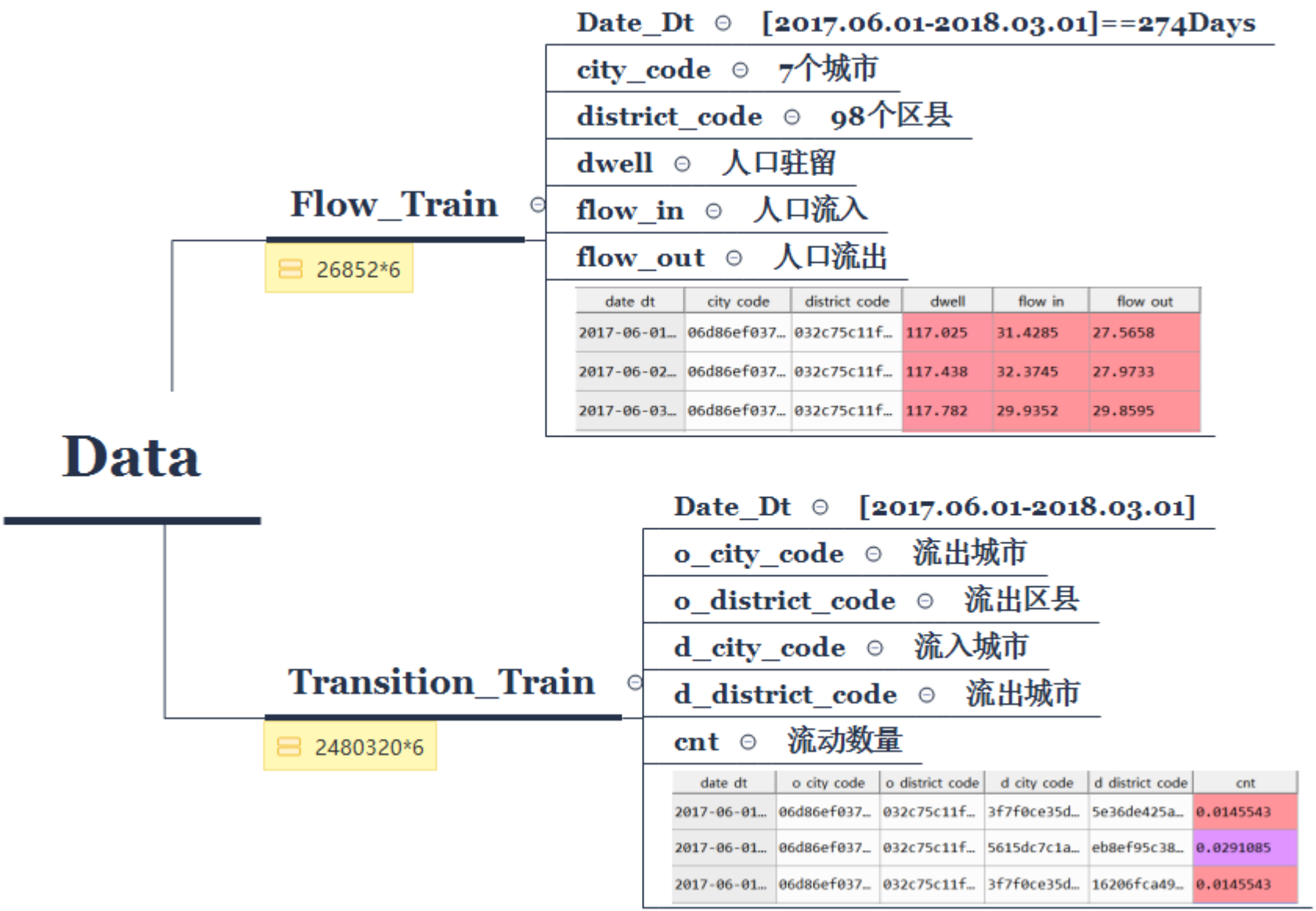
报告内容

目标

- 利用几个邻近城市的移动通信设备用户数历史变动情况，各区县之间的用户转移情况等数据，，对上述城市各个区县未来 15 天的总人口变化情况进行动态预测
- 预测每个地区未来 15 天每天的人口临时驻留，人口流入，人口流出
- 损失评估，其中 T 为日期，N 为地区，k 为驻留，流入，流出：

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{k=1}^{3}\left(\log\left(\hat{y}_{tnk}+1\right)-\log\left(y_{tnk}+1\right)\right)^2}{T*N*3}}$$

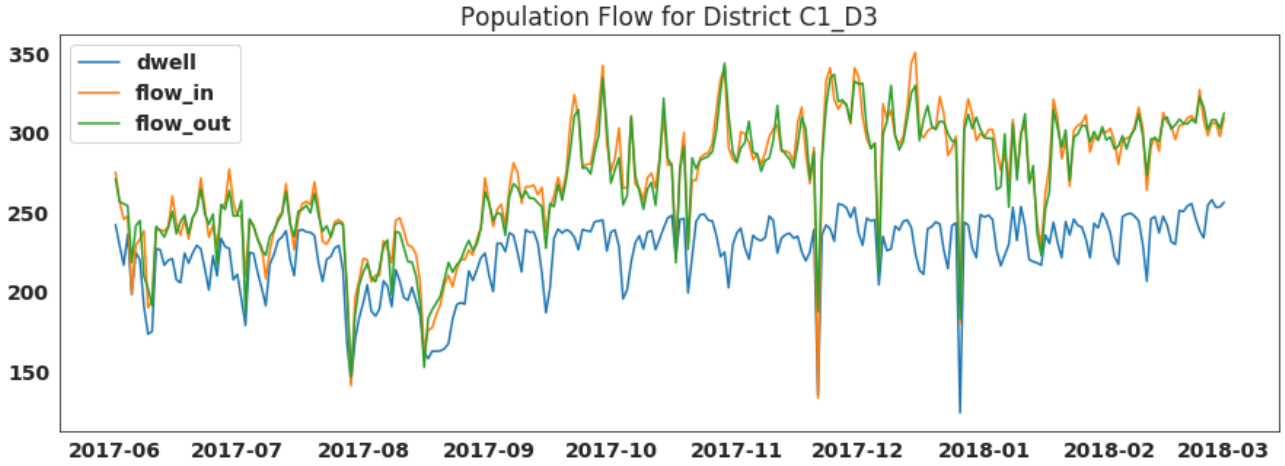
数据

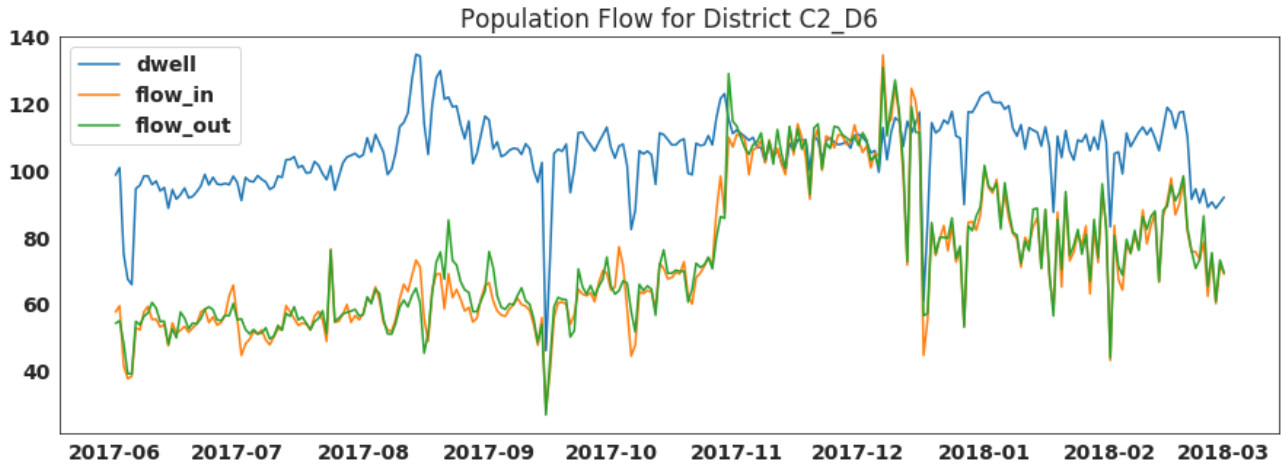


date dt	city code	district code	dwell	flow in	flow out
2017-06-01…	06d86ef037…	032c75c11f…	117.025	31.4285	27.5658
2017-06-02…	06d86ef037…	032c75c11f…	117.438	32.3745	27.9733
2017-06-03…	06d86ef037…	032c75c11f…	117.782	29.9352	29.8595

date dt	o city code	o district code	d city code	d district code	cnt
2017-06-01…	06d86ef037…	032c75c11f…	3f7f0ce35d…	5e36de425a…	0.0145543
2017-06-01…	06d86ef037…	032c75c11f…	5615dc7c1a…	eb8ef95c38…	0.0291085
2017-06-01…	06d86ef037…	032c75c11f…	3f7f0ce35d…	16206fca49…	0.0145543

数据特点





- 区县间差异大

  解决方案

  ⇒ 区县单独建模

  ⇒ 城市建模

  ⇒ 关联性建模

- 假期阶段性异常

  解决方案

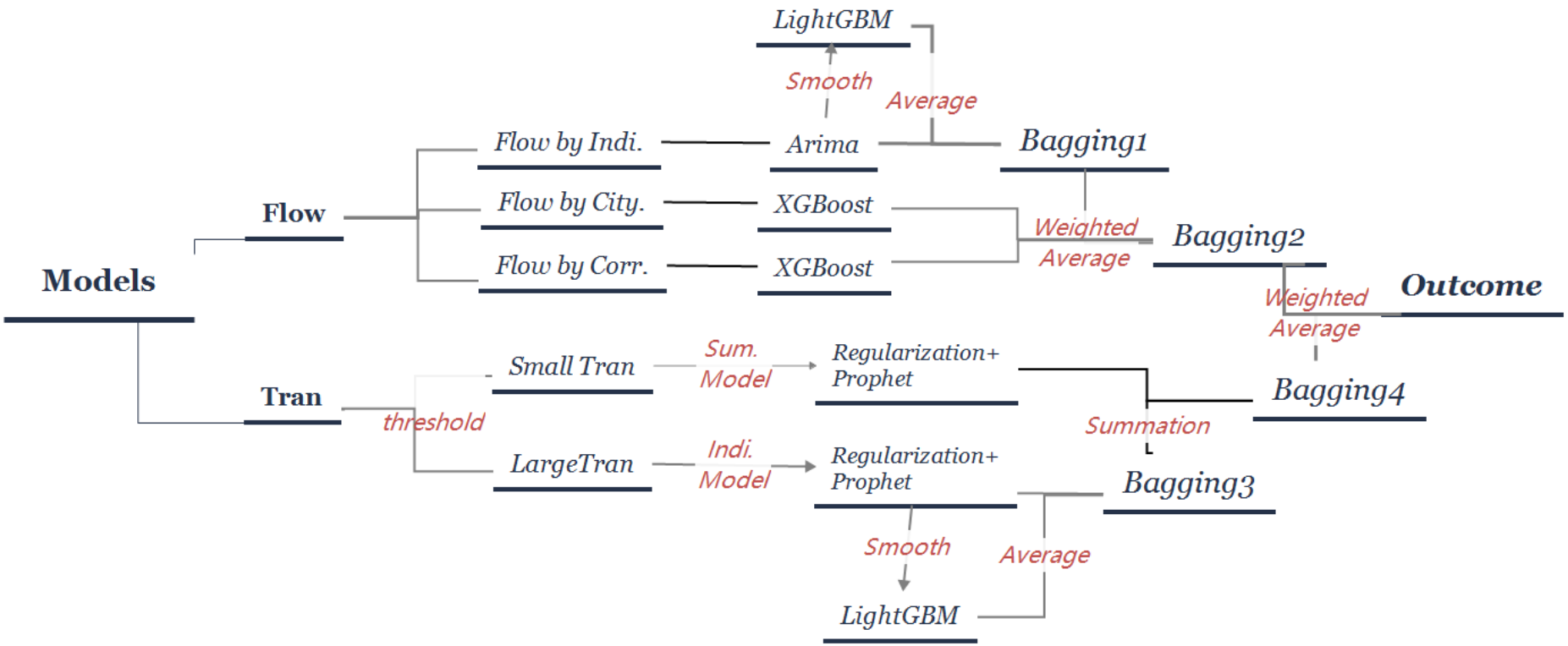
  ⇒ 检测区间并删除

- 异常点较多

  解决方案

  ⇒ 检测异常并平滑

模型方法

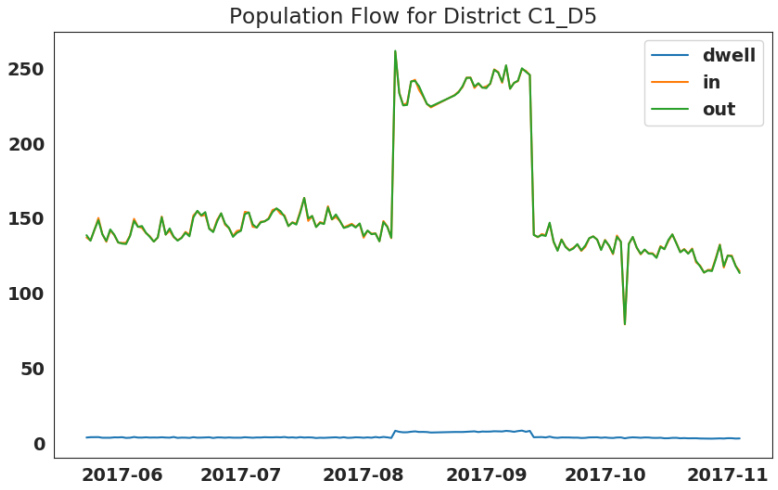

LightGBM
Smooth
Average
Flow by Indi.
Arima
Bagging1
Flow
Flow by City.
XGBoost
Weighted Average
Bagging2
Models
Flow by Corr.
XGBoost
Weighted Average
Outcome
Small Tran
Sum. Model
Regularization+ Prophet
Tran
Bagging4
threshold
Summation
LargeTran
Indi. Model
Regularization+ Prophet
Bagging3
Smooth
Average
LightGBM

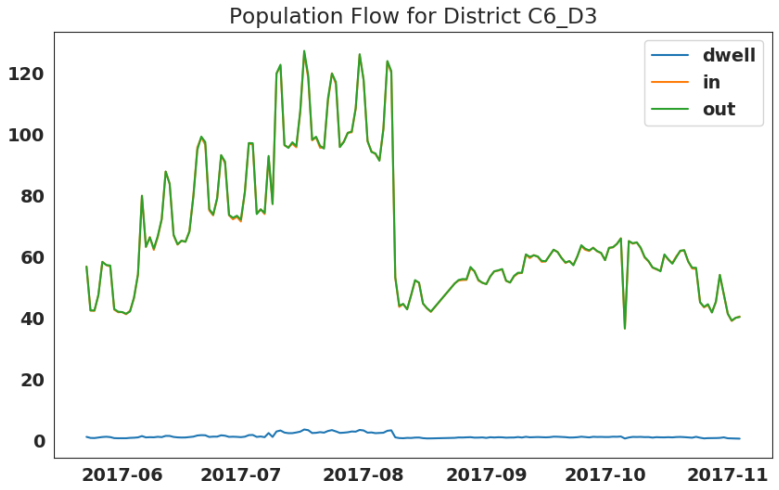
报告内容

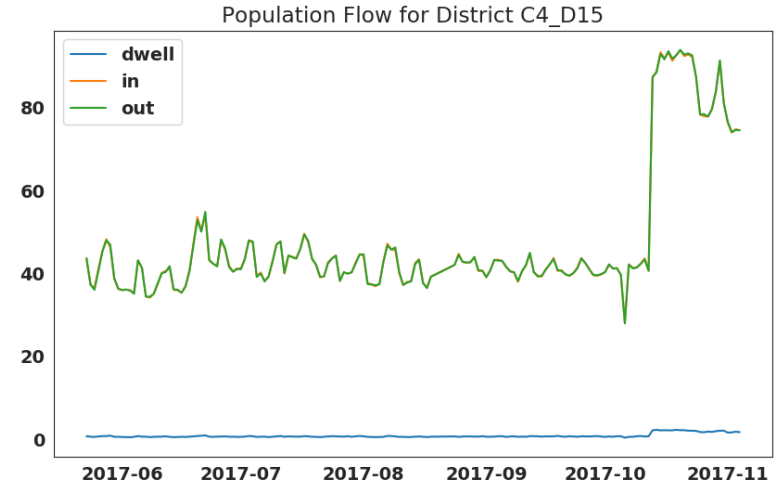
目标

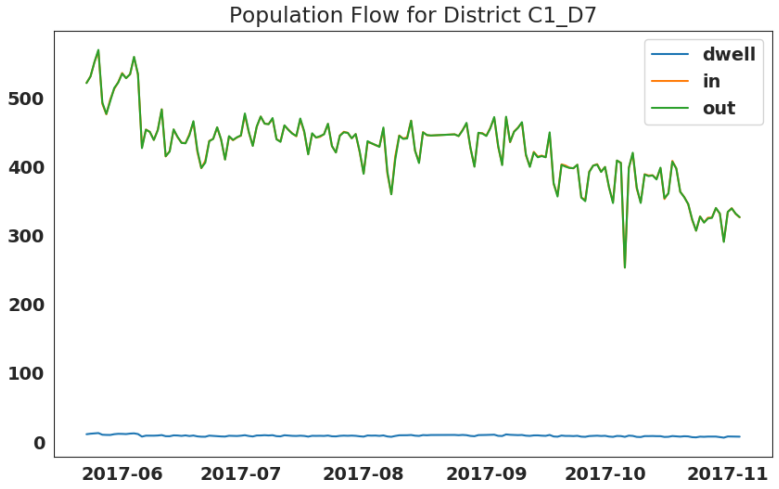
- 更换新的数据集，数据来源、时间地点均有所变动，数据格式和初赛阶段一致，但在训练数据集中去掉了中间 5 天的数据，提交这 5 天的缺失数据值，同时预测未来 10 天的数据。
- 数据量变大，由 7 个城市变 13 个城市，区县由 98 变为 204 个城市，但每个区县的时间长度缩短，由 $[2017.06.01, 2018.03.01]$ 缩短为 $[2017.05.23, 2017.11.04]$.
- $2017.08.19 - 2017.08.23$ 五天缺失 $+$ $2017.11.05 - 2017.11.15$ 未来预测.
- 损失函数与相同.

数据特征及处理


Population Flow for District C1_D5
dwell
in
out
250
200
150
100
50
0
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11


Population Flow for District C6_D3
dwell
in
out
120
100
80
60
40
20
0
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11


Population Flow for District C4_D15
dwell
in
out
80
60
40
20
0
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11


Population Flow for District C1_D7
dwell
in
out
500
400
300
200
100
0
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11

解决思路

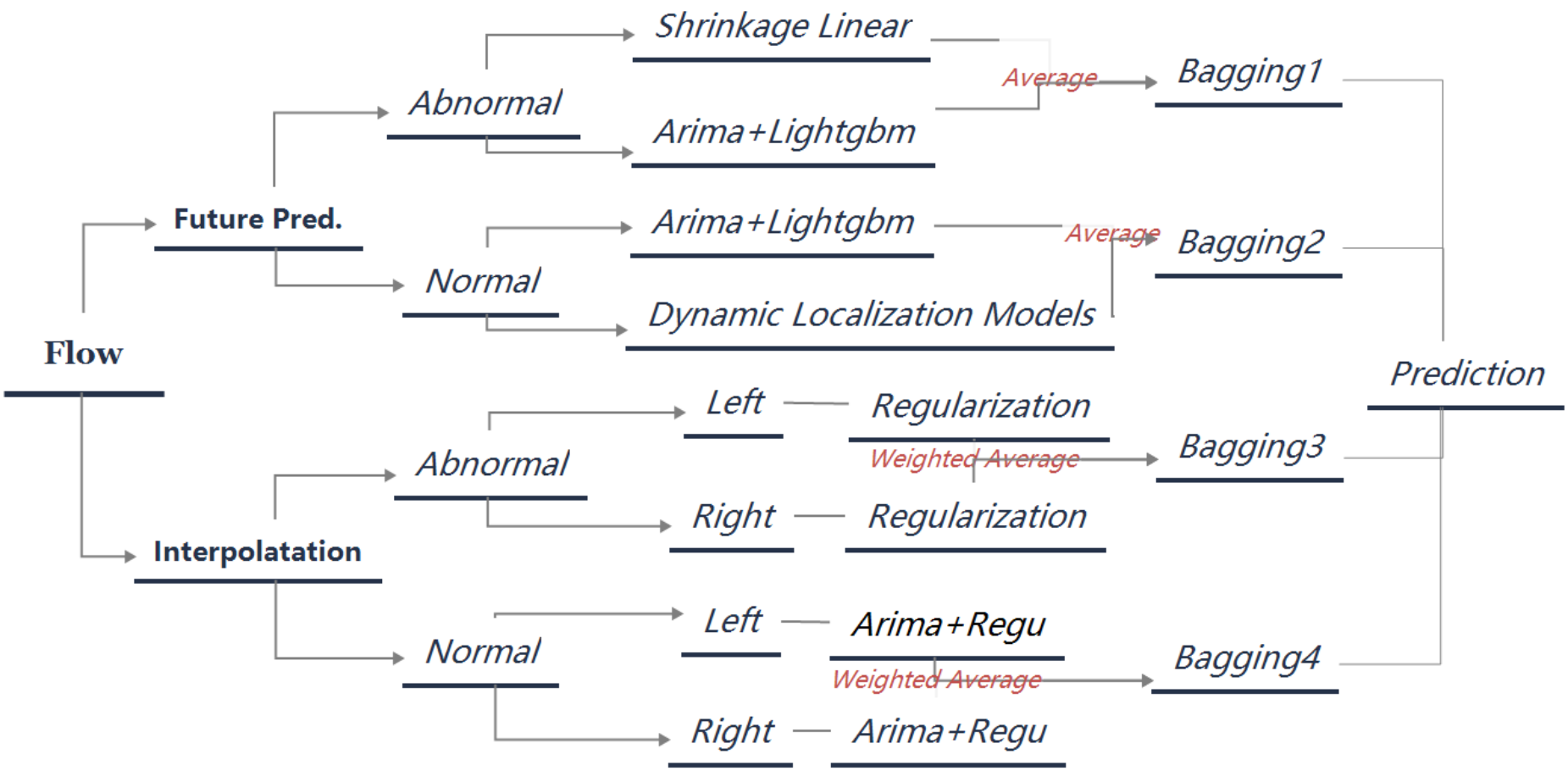
未来预测

- 先删除缺失值左右各一天数据，保留原始数据周的周期性;
- 对于位于中间的异常区间，检测并删除，删除区间以周为单位;
- 但若异常区间位于数据末端，则单独进行建模;
- 对于异常点进行平滑;

中间缺失插补

- 若 08.19-08.23 位于异常时段，则将异常时段单独提取出来简单建模;
- 其他进行单独建模;
- 建模思路：将数据提取出来，分左右两端，分别进行建模，得到缺失预测值，然后将两端进行加权平均。

模型方法


Flow
Future Pred.
Abnormal
Shrinkage Linear
Arima+Lightgbm
Average
Bagging1
Normal
Arima+Lightgbm
Dynamic Localization Models
Average
Bagging2
Interpolatation
Abnormal
Left
Regularization
Right
Regularization
Weighted Average
Bagging3
Normal
Left
Arima+Regu
Right
Arima+Regu
Weighted Average
Bagging4
Prediction

报告内容

Any Ideas !

Welcome !

Thank You !