ETP Course

Practice of AI

Time Series Forecasting

Jim Xie

2020/7/6

Demo

Forecasting for the count infected by COVID-19

(USA)

https://www.kaggle.com/sudalairajkumar/covid19-in-usa

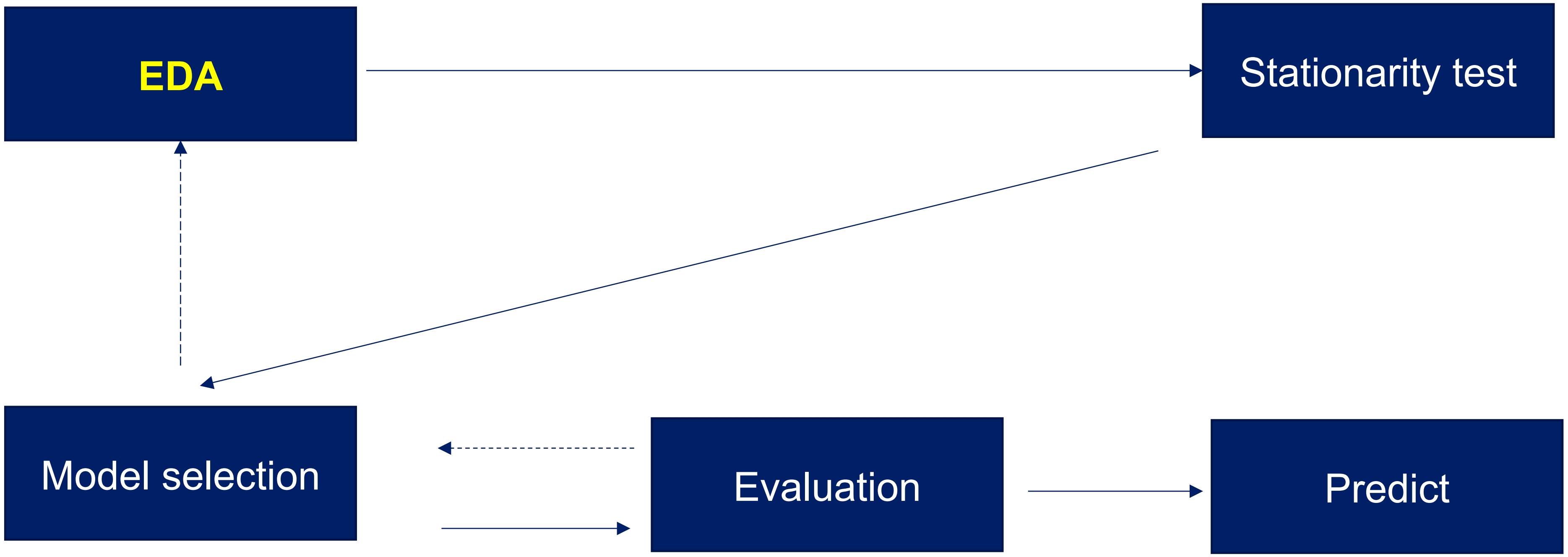
Sample

Dataset： 147 (2020-01-22 to 2020-06-16)
Dimension： 25

- Date
- PositiveIncrease
- States
- TotalTestResults

```
Out[6]: date                          object
        states                        int64
        positive                      int64
        negative                      float64
        pending                       float64
        hospitalizedCurrently         float64
        hospitalizedCumulative        float64
        inIcuCurrently                float64
        inIcuCumulative               float64
        onVentilatorCurrently         float64
        onVentilatorCumulative        float64
        recovered                     float64
        dateChecked                   object
        death                         float64
        hospitalized                  float64
        lastModified                  object
        total                         int64
        totalTestResults              int64
        posNeg                        int64
        deathIncrease                 int64
        hospitalizedIncrease          int64
        negativeIncrease              int64
        positiveIncrease              int64
        totalTestResultsIncrease      int64
        hash                          object
```

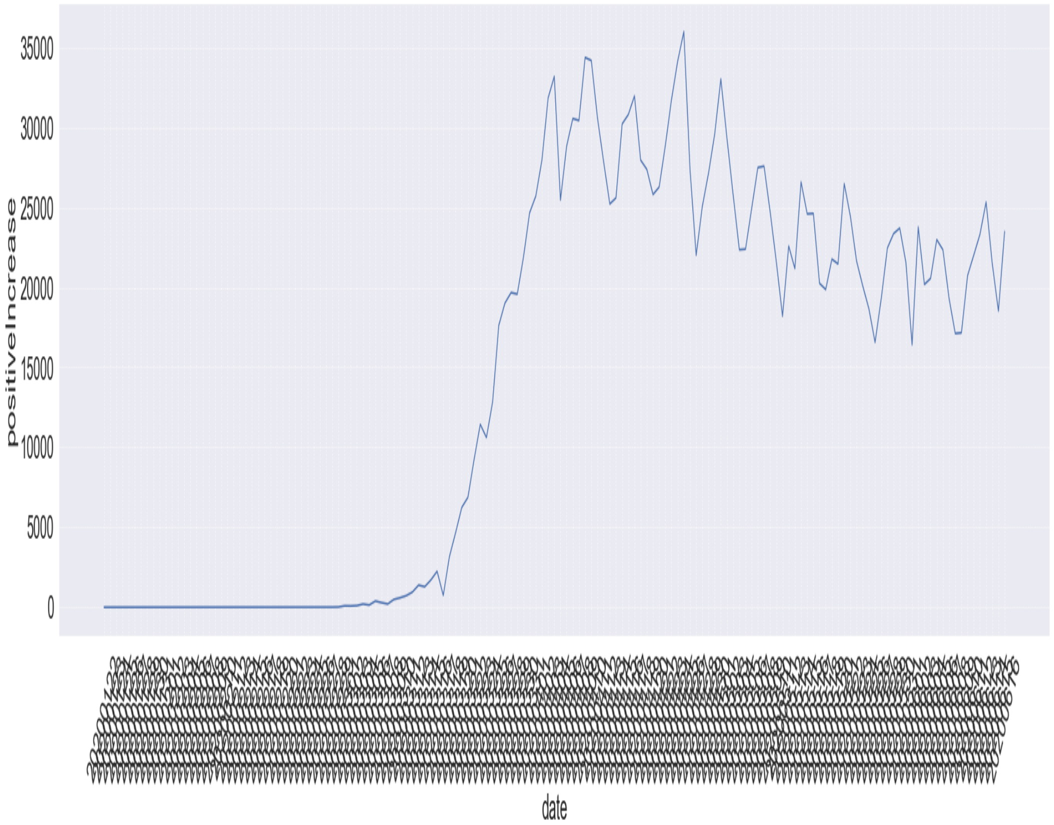
Workflow



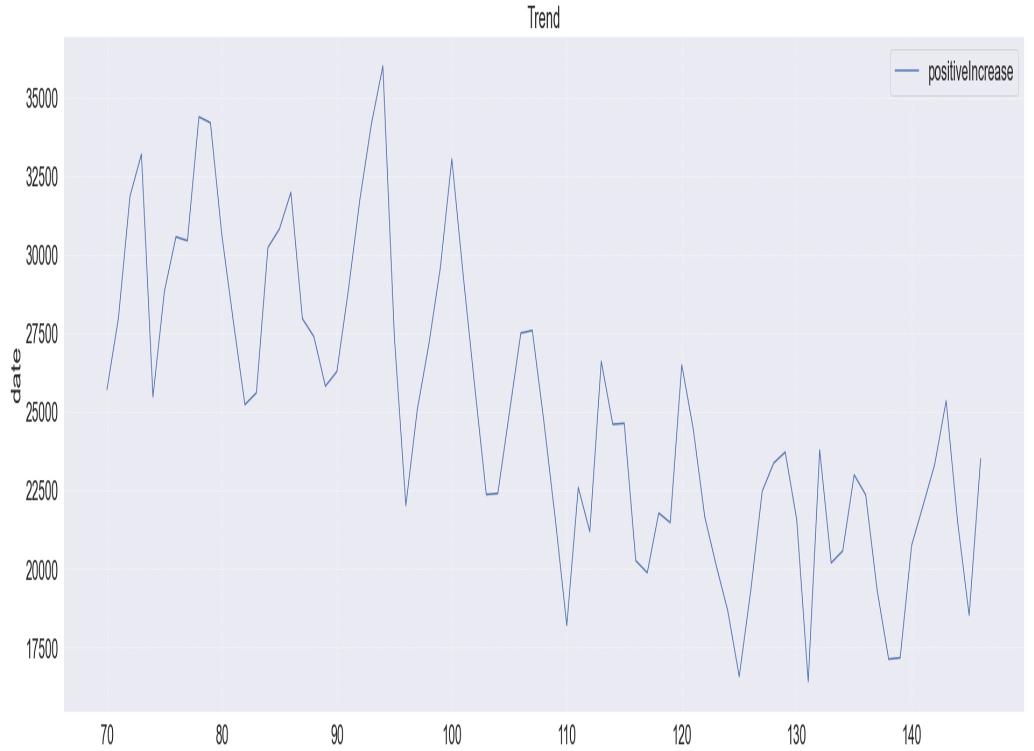
EDA
Stationarity test
Model selection
Evaluation
Predict

EDA # Positive Increase

- Increase trend

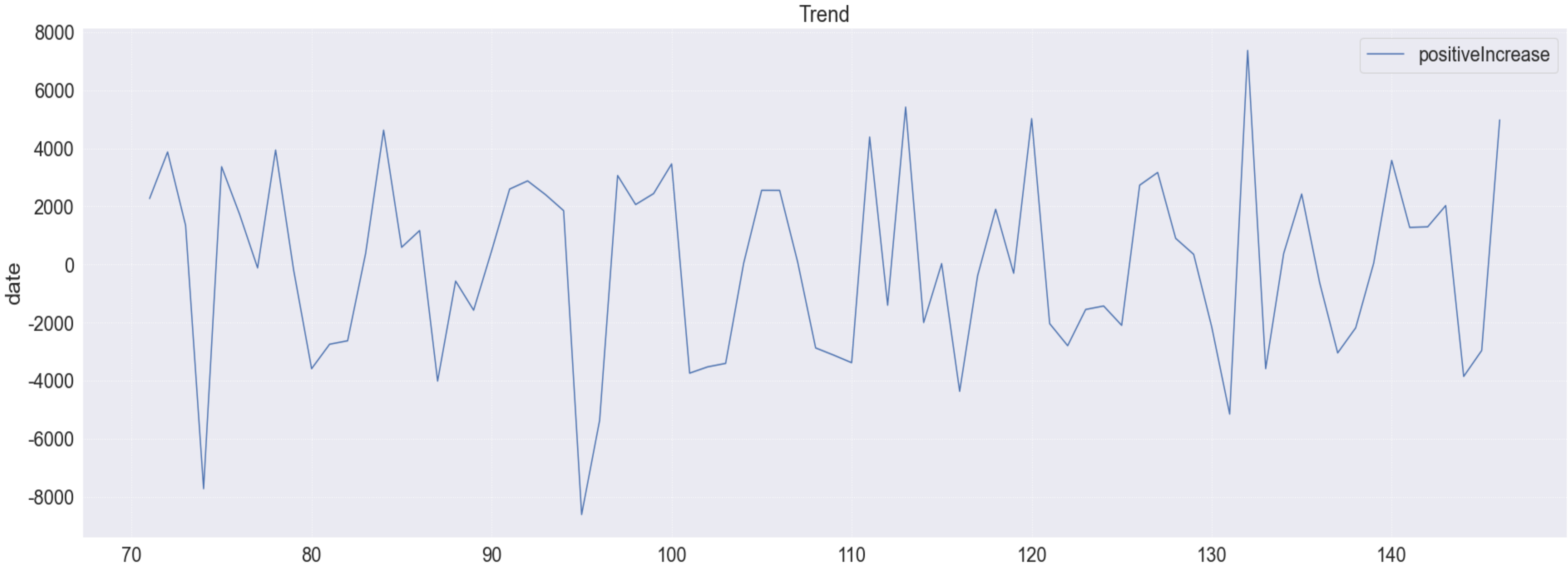


- Decline convulsively after 4/1



EDA # Positive Increase

❖ Stable after difference



EDA # Positive Increase & Test

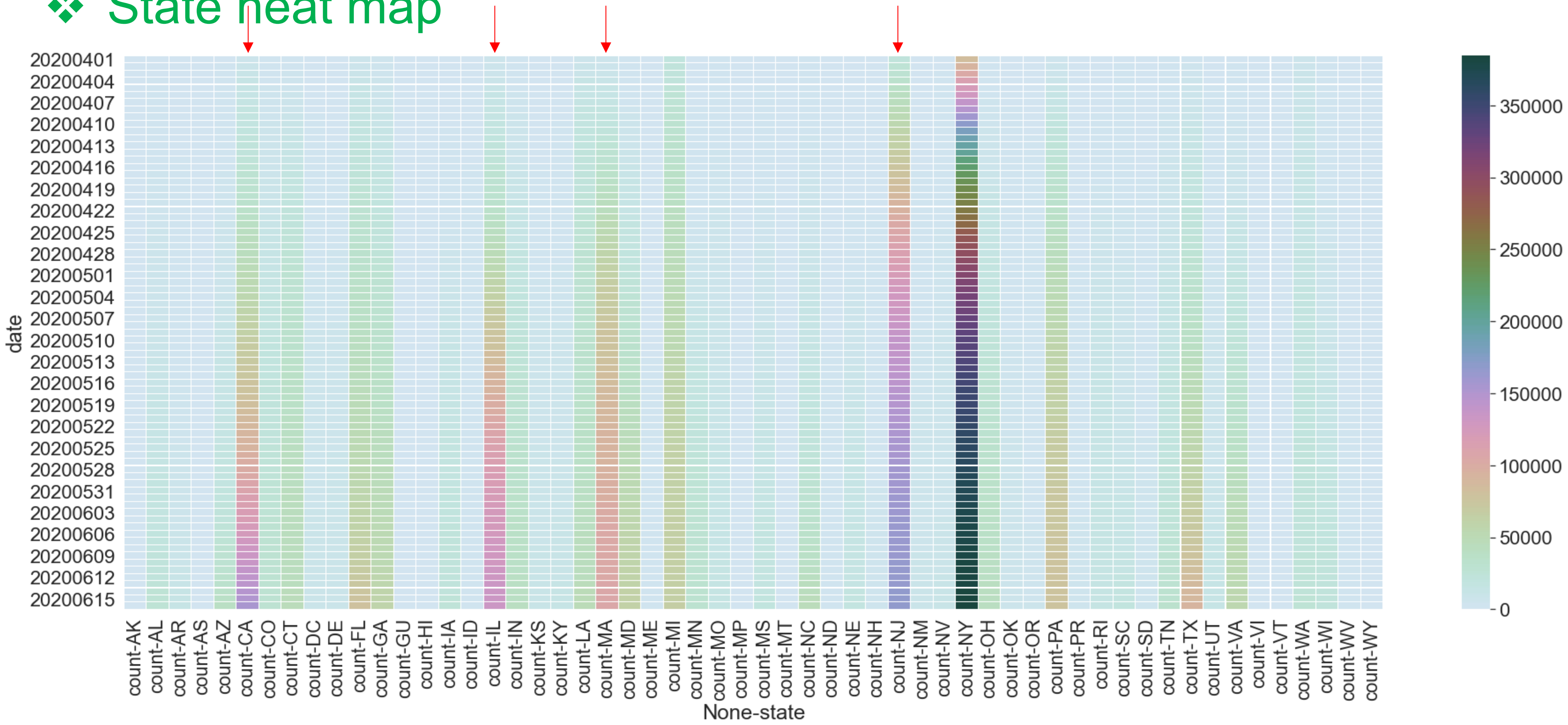
❖ Correlation matrix

:

	positive	positiveIncrease	test	testIncrease
positive	1.00	-0.74	0.97	0.94
positiveIncrease	-0.74	1.00	-0.71	-0.62
test	0.97	-0.71	1.00	0.93
testIncrease	0.94	-0.62	0.93	1.00

EDA # Positive Count & State

❖ State heat map

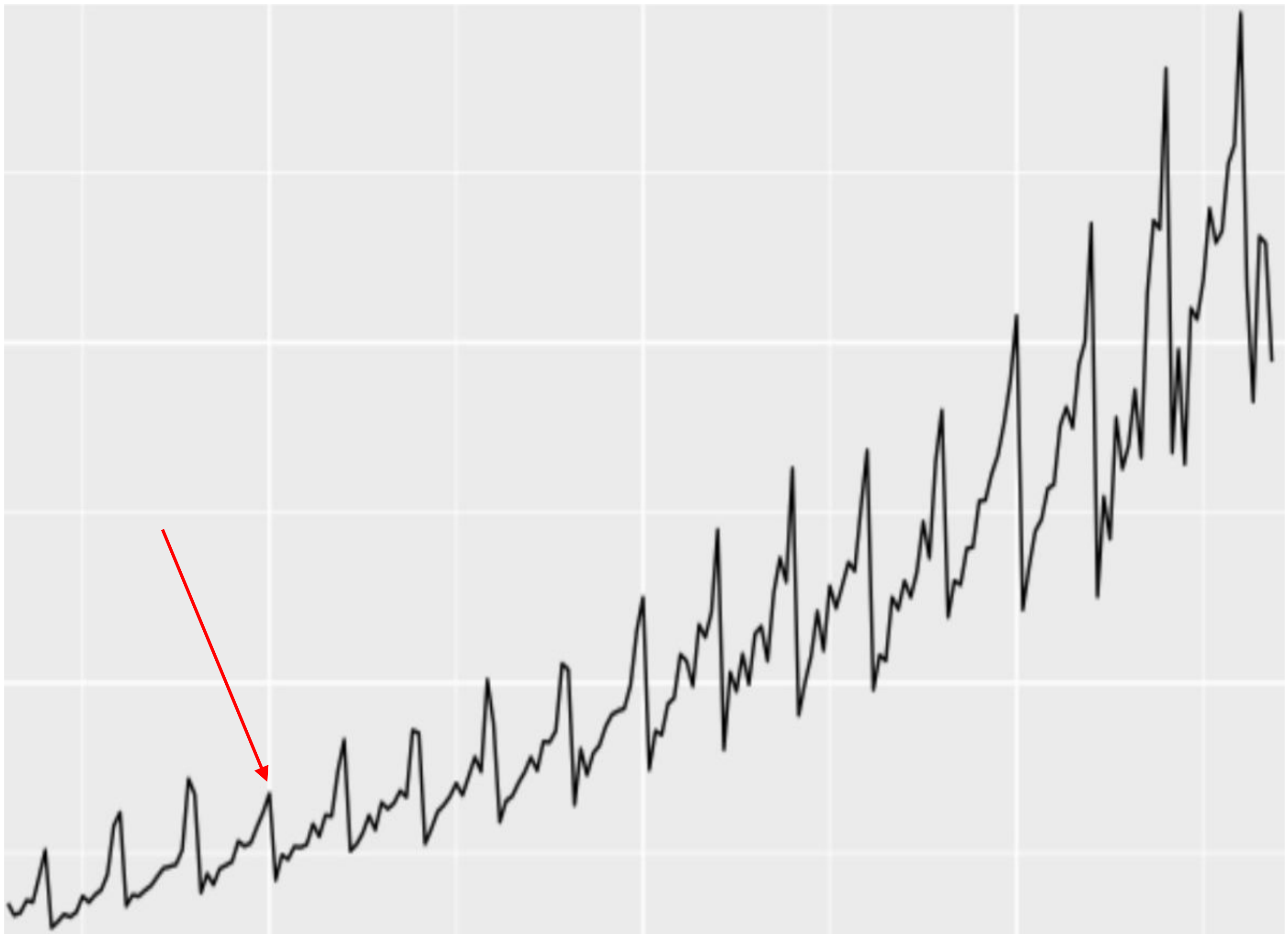


EDA

- ✓ Increase declines convulsively after 4/1
- ✓ Increase becomes stable after difference
- ✓ High correlation between increase and test
- ✓ No increase pattern found in different states

Time Series Data

- Trend
- Seasonal
- Cyclical
- Irregular

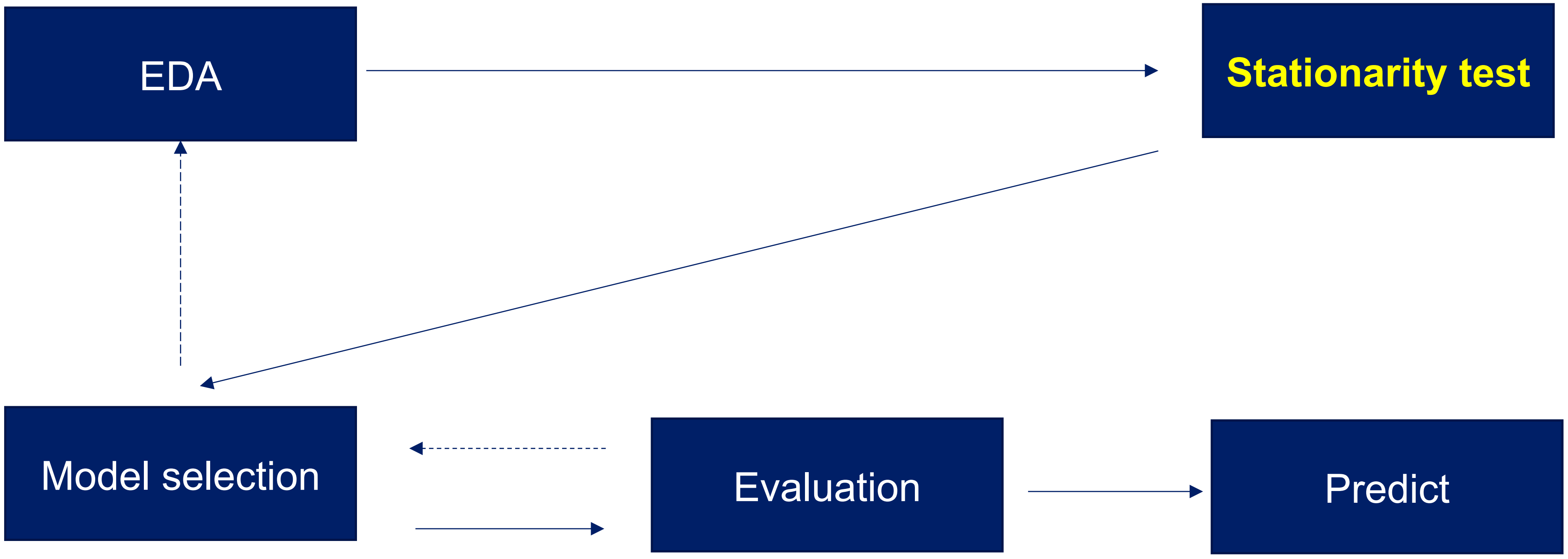


Model

- **ARIMA** $$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

AR	+	I	+	MA
Autoregressive		Integrated		Moving Average
$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t$		Difference		$y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$
P		D		Q

Workflow


EDA
Stationarity test
Model selection
Evaluation
Predict

Stationarity Test

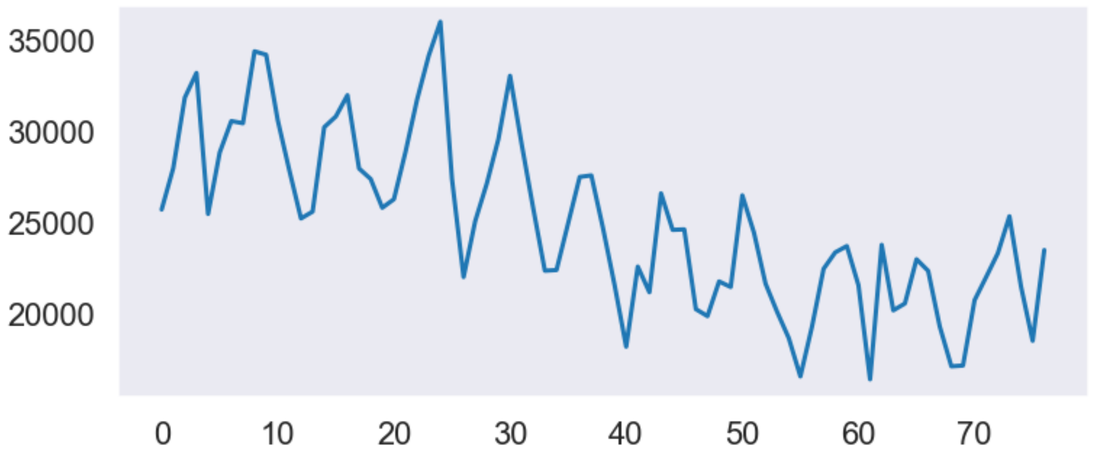
❖ ADF test result

- ✓ P Value < 0.05
- ✓ Test Statistic < Critical Value

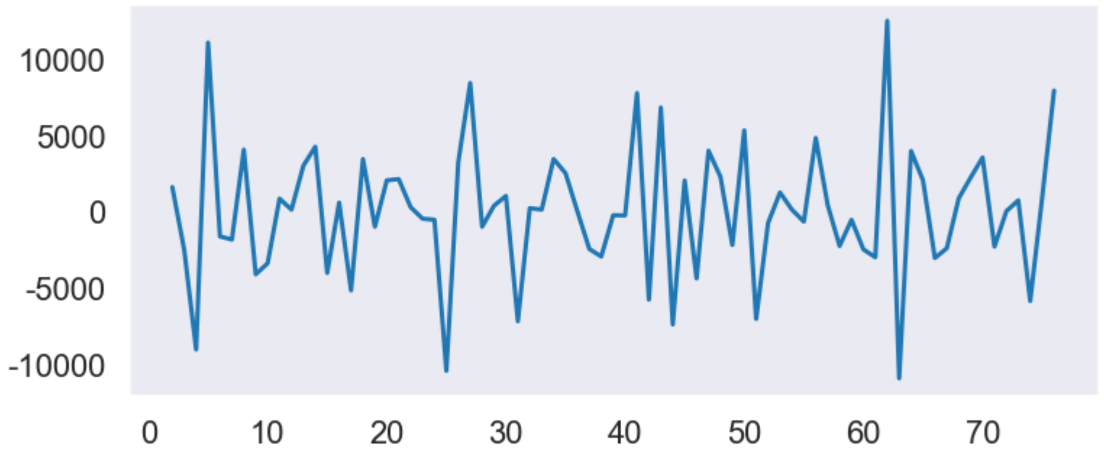
Stage	Test result
Raw data	Test Statistic -0.94 p-value 0.78 Critical Value (1%) -3.53 Critical Value (5%) -2.90 Critical Value (10%) -2.59
After smooth	Test Statistic -1.20 p-value 0.68 Critical Value (1%) -3.54 Critical Value (5%) -2.91 Critical Value (10%) -2.59
After difference 1	Test Statistic -9.86 p-value 0.00 Critical Value (1%) -3.53 Critical Value (5%) -2.90 Critical Value (10%) -2.59
After difference 2	Test Statistic -8.07 p-value 0.00 Critical Value (1%) -3.53 Critical Value (5%) -2.91 Critical Value (10%) -2.59

Stationarity

- Stable after difference 2



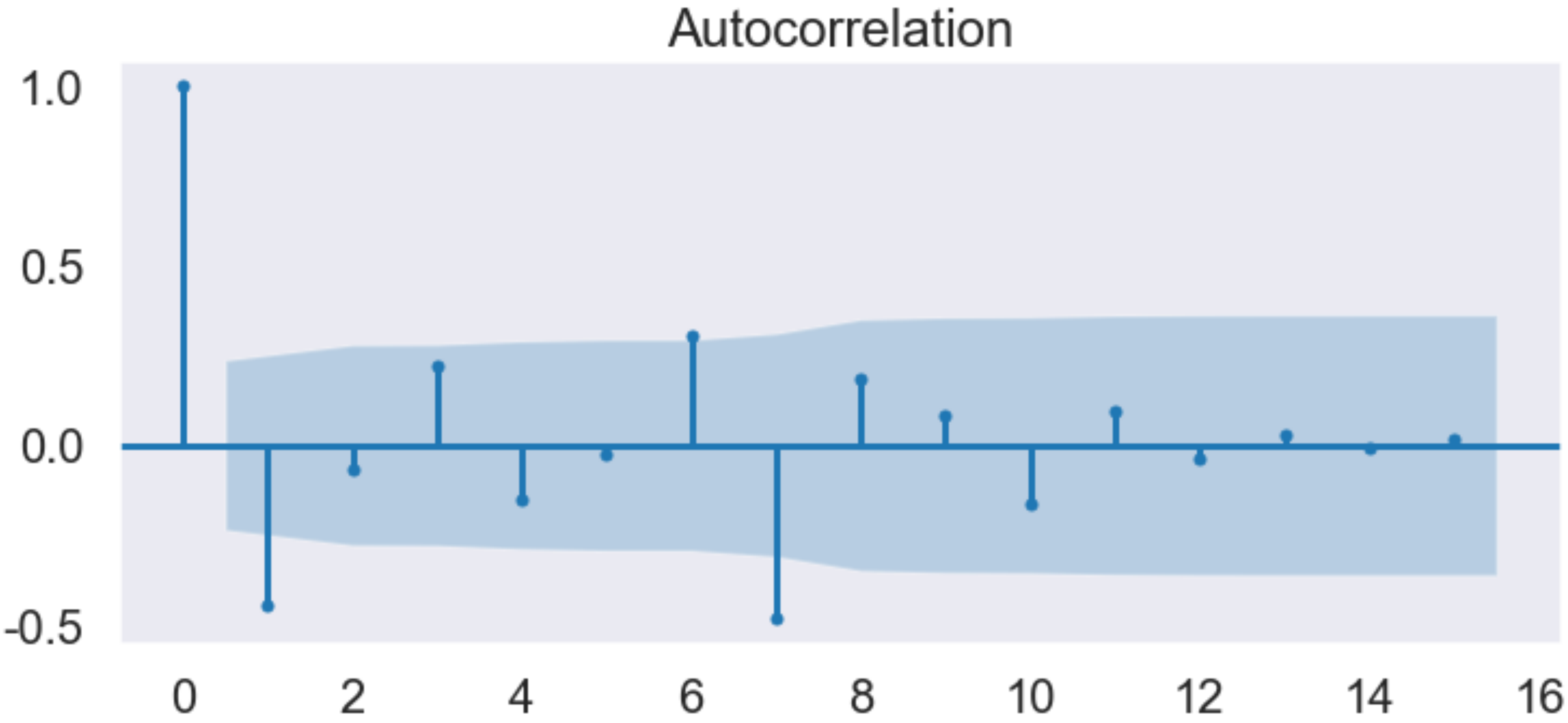
Raw data



After difference 2

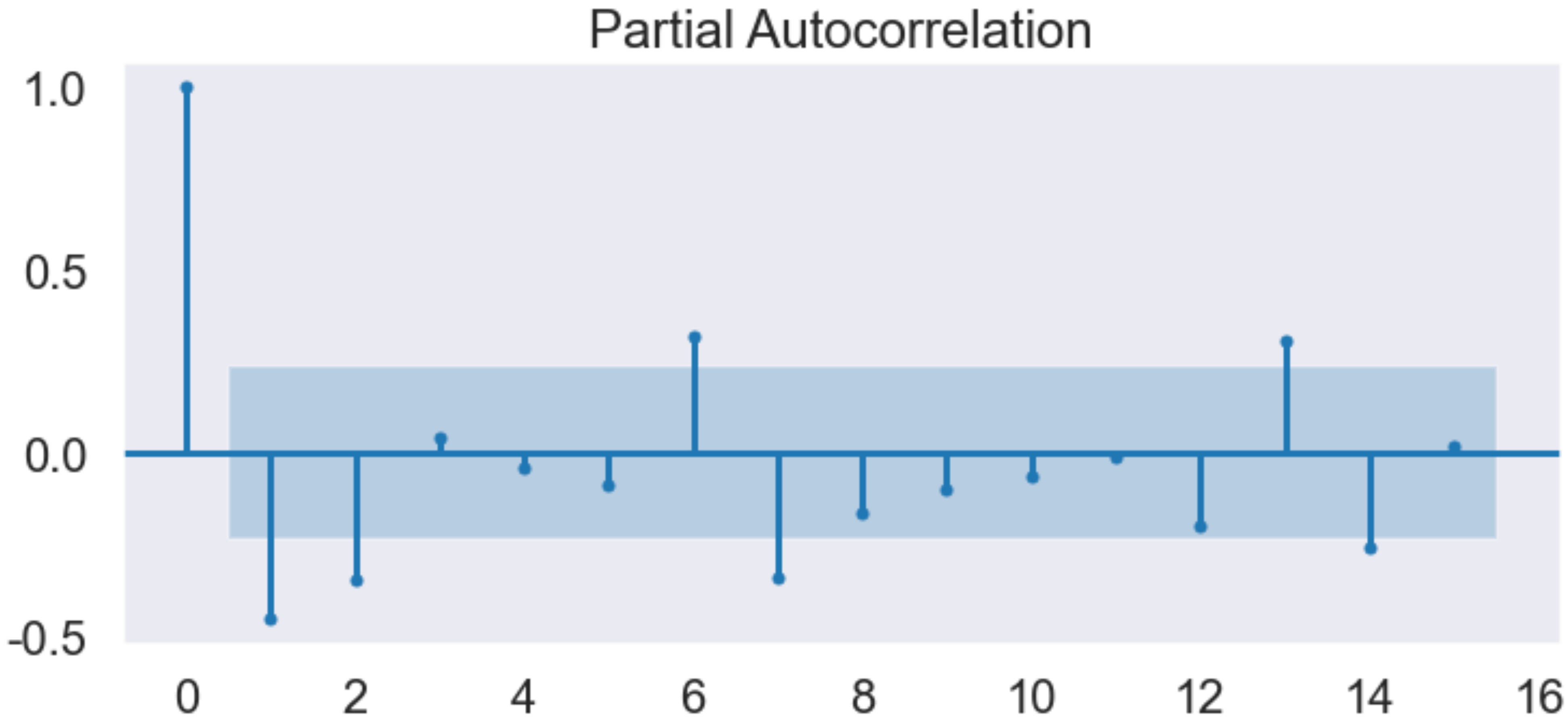
ACF

- Q = 2



PACF

- P = 3



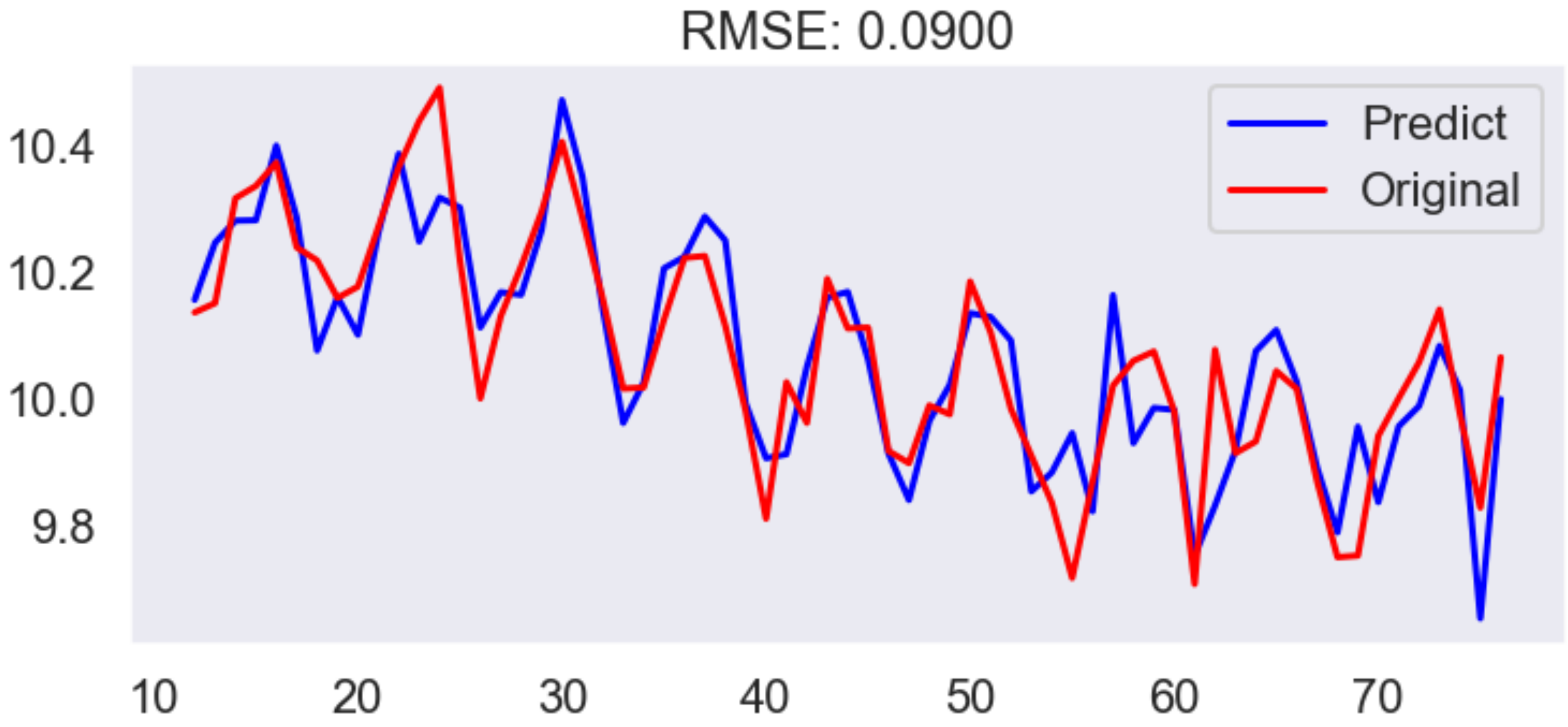
Model/P/Q

Model	AR(p)	MA(q)	ARMA(p,q)
ACF	Tails off	Truncated after N Lag	Tails off
PACF	Truncated after N Lag	Tails off	Tails off

- Recommended parameter pair (P=3,D=2,Q =2)
- Try parameter pair (p in [0:5],q in [0:5]

Verify

❖ $RMSE(X,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x_i)-y_i)^2}$



Best performance (P=4,Q=4)

P=1,Q=0,Error=3044.832934
P=1,Q=1,Error=2945.576680
P=1,Q=2,Error=2966.281480
P=1,Q=3,Error=2896.684189
P=1,Q=4,Error=2860.274045
P=1,Q=5,Error=2788.130198
P=2,Q=0,Error=2855.045201
P=2,Q=1,Error=2839.954995
P=2,Q=2,Error=2741.163190
P=2,Q=3,Error=2638.373490
P=2,Q=5,Error=2861.528584
P=3,Q=0,Error=2710.645780
P=3,Q=1,Error=2704.595950
P=3,Q=2,Error=2611.901070
P=3,Q=3,Error=2428.514544
………

Forecast

❖Predict

Date	Real	Predicted	Error	E(%)
20200621	27287	21909	5378	19.7090
20200620	31958	21816	10142	31.7354
20200619	31055	22319	8736	28.1307
20200618	27512	23633	3879	14.0993
20200617	**23871**	**23407**	**464**	**1.9438**

Further

- Samples
  1. More verify samples
  2. Decompose
  3. Infected count explode
  4. …………..

- More features
  1. Medical information
  2. Population density
  3. Seasonality
  4. …………..

- More models
  1. RNN
  2. LSTM
  3. Model with CNN
  4. …………..

Demo

Full flow

Learn from demo

- ✓ More effort in EDA and FE
- ✓ Model has pre-condition and limitation
- ✓ Try several models or parameter pairs
- ✓ Evaluate and select suitable model

Thanks

2020-8-15

❖ Backlog

Backlog

❖ 单位根

DF检验

假设Y是由一阶自回归过程生成

$$Y_t = \gamma Y_{t-1} + v_t \quad (35)$$

若 $|\gamma| < 1$，则Y是平稳的,若 $|\gamma| = 1$，则Y是非平稳的

检验问题： $H_0: \gamma = 1, \quad H_A: \gamma < 1$

模型转换之后 $$\Delta Y_t = (\gamma - 1)Y_{t-1} + v_t = \beta Y_{t-1} + v_t \quad (36)$$

检验问题： $H_0: \beta = 0, \quad H_A: \beta < 1$

❖ 单位根&白噪音检验

单位根检验

方法：statsmodels.tsa.stattools import adfuller

判断：

1. 结果同时小于1%、5%、10%对应的值，表示平稳
2. P-value (不变显著性) 接近0 ，表示平稳

白噪音检验

方法: statsmodels.stats.diagnostic import acorr_ljungbox

判断：

1. P-value < 0.05 表示不是白噪音