

1.只有 PM2.5 的時間序列迴歸模型

使用前六期的 PM2.5 資料預測下一期的 PM2.5

迴歸方程式如下：

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 Y_{t-4} + \beta_5 Y_{t-5} + \beta_6 Y_{t-6} + \epsilon_t$$

變數名稱	係數
截距項	10.629672551778064
LAG1	0.14592256
LAG2	0.03935985
LAG3	0.04966078
LAG4	-0.04638318
LAG5	0.0303036
LAG6	0.09881222

2.只有 PM2.5 的隨機森林

參數設定:500 棵樹、leaf node 至少包含五筆資料、每次隨機選取 2 個特徵值做 best split(因為一共有 6 個落後期的資料，因為是迴歸問題，因此最好每次隨機篩選 $p/3=6/3=2$ (round down)個特徵值)，並且設定隨機種子為 0

3.所有 18 種屬性的時間序列迴歸模型

使用前六期的資料去預測下一期的 PM2.5，其中前六期的資料包含所有的觀測項目，因此一共有 $18*6$ 個特徵。

$$Y_t = \beta_0 + \sum_{\substack{i \in \text{測項} \\ j=1}}^6 \beta_{i,t-j} Y_{i,t-j} + \epsilon_t$$

以下截圖是它的迴歸係數(從第一個屬性的 LAG1 開始，接續是第一個屬性的 LAG2...)

截距項: -88.7364257352892

```

In [2]: reg1.coef_
Out[2]:
array([ 8.42177067e-01, -4.69764335e-01, -2.25595102e-01,  6.75909414e-02,
        -6.06313170e-02, -2.68155442e-01,  6.92478994e+00, -7.13687525e+00,
        -6.78824166e+00,  1.42678649e+01,  5.17120770e+00,  1.70475173e+01,
        -8.02064077e+00,  2.30551908e+00,  4.26849153e+00, -1.03370069e+01,
        6.98014584e+00,  7.39629878e+00,  2.93734034e+00,  9.08215570e+00,
        -4.40029102e+00,  2.78125012e+01, -3.36180588e+01, -3.34662533e+01,
        -6.63258652e-01, -1.04427095e+00, -4.83712278e-01,  5.89927950e-01,
        -2.55948127e-01,  9.48201695e-01, -4.74076869e-01, -7.15685956e-01,
        -2.64986301e-01,  7.66663420e-01, -3.22623302e-01,  1.13135953e+00,
        6.31694675e-01,  9.28998712e-01,  3.94591407e-01, -4.13111885e-01,
        2.82527077e-01, -1.09102313e+00,  1.39836030e-01,  1.04497016e-01,
        1.78989326e-02,  9.67106703e-02,  1.73740900e-02,  3.67388896e-02,
        6.12907567e-02,  2.60812147e-02, -9.20661027e-02, -1.10067894e-01,
        5.68311115e-02,  1.14288748e-01, -2.12785706e-01,  1.50474857e-02,
        -6.58045237e-02,  2.57156121e-02, -8.81423906e-02, -1.51620508e-01,
        -2.80534687e-02, -9.04623061e-01, -4.72617289e-01, -7.77034714e-01,
        -1.23035699e+00, -8.72125898e-01,  5.37270538e-02, -1.48049898e-01,
        1.43528169e-01,  2.00000783e-01, -3.61121704e-03,  1.52222729e-01,
        1.12846359e-01, -1.18374809e+00,  8.03981672e-01, -1.15600004e+00,
        1.57139610e-01, -9.11753749e-02, -4.14799925e+00, -1.14948523e-01,
        4.22222120e+00, -1.54873156e+01,  1.06323940e+01,  6.03535682e+00,
        2.25846199e-03,  2.71136792e-03,  7.74209510e-03,  1.70464436e-03,
        2.01005338e-03, -1.45811956e-03,  2.09178540e-02, -1.24348271e-02,
        3.05465614e-05,  4.07675847e-03, -7.34612387e-03, -2.47278070e-03,
        -5.88779739e-02,  1.13456650e+00,  3.18365759e-01, -1.89216384e-01,
        9.47238120e-01, -9.83999937e-01, -1.08051249e+00,  4.31680955e-02,
        6.79681356e-01,  6.81952330e-01,  1.06320933e-01,  8.99380761e-01])

```

4.所有 18 種屬性的隨機森林

參數設定:500 棵樹、leaf node 至少包含五筆資料、每次隨機選取 2 個特徵值做 best split(因為一共有 6 個落後期的資料，因為是迴歸問題，因此最好每次隨機篩選 $p/3=6/3=2$ (round down)個特徵值)，並且設定隨機種子為 0

☆ 模型績效(MAE) $\frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$ 用測試資料與四種模型預測的結果作比對

MAE	
只有 PM2.5 的時間序列迴歸模型	9.936985320586759*
只有 PM2.5 的隨機森林	9.975633760850334
所有 18 種屬性的時間序列迴歸模型	11.624869392277475
所有 18 種屬性的隨機森林	9.777989700481417

由結果我們可以看出，只有利用 PM2.5 的屬性的前六期資料去預測當期 PM2.5 的時間序列模型，若以 MAE 來衡量模型的表現，它是表現得最好的。