資料探勘於實際數據之應用實作

製作人:楊翔斌

大綱

第一部分：資料探勘應用在類別變數預測-以鋼板缺陷類型為例
第二部分：資料探勘應用在連續變數預測-以台北市房屋價格為例
第三部分：利用CNN做圖片類別辨識別辨識
第四部分：利用LSTM做時間序列相關預測
第五部分：資料視覺化
第六部分：應用文字探勘於文章分類
第七部分：用決策樹作是否離職之分類

第一部分：資料內容

```
Classes 'data.table' and 'data.frame':  1941 obs. of  35 variables:
 $ X_Minimum            : int  42 645 829 853 1289 430 413 190 330 74 ...
 $ X_Maximum            : int  50 651 835 860 1306 441 446 200 343 90 ...
 $ Y_Minimum            : int  270900 2538079 1553913 369370 498078 100250 138468 210936 429227 779144 ...
 $ Y_Maximum            : int  270944 2538108 1553931 369415 498335 100337 138883 210956 429253 779308 ...
 $ Pixels_Areas         : int  267 108 71 176 2409 630 9052 132 264 1506 ...
 $ X_Perimeter          : int  17 10 8 13 60 20 230 11 15 46 ...
 $ Y_Perimeter          : int  44 30 19 45 260 87 432 20 26 167 ...
 $ Sum_of_Luminosity    : int  24220 11397 7972 18996 246930 62357 1481991 20007 29748 180215 ...
 $ Minimum_of_Luminosity: int  76 84 99 99 37 64 23 124 53 53 ...
 $ Maximum_of_Luminosity: int  108 123 125 126 126 127 199 172 148 143 ...
 $ Length_of_Conveyer   : int  1687 1687 1623 1353 1353 1387 1687 1687 1687 1687 ...
 $ TypeOfSteel_A300     : int  1 1 1 0 0 0 0 0 0 0 ...
 $ TypeOfSteel_A400     : int  0 0 0 1 1 1 1 1 1 1 ...
 $ Steel_Plate_Thickness: int  80 80 100 290 185 40 150 150 150 150 ...
 $ Edges_Index          : num  0.0498 0.7647 0.971 0.7287 0.0695 ...
 $ Empty_Index          : num  0.241 0.379 0.343 0.441 0.449 ...
 $ Square_Index         : num  0.1818 0.2069 0.3333 0.1556 0.0662 ...
 $ Outside_X_Index      : num  0.0047 0.0036 0.0037 0.0052 0.0126 0.0079 0.0196 0.0059 0.0077 0.0095 ...
 $ Edges_X_Index        : num  0.471 0.6 0.75 0.538 0.283 ...
 $ Edges_Y_Index        : num  1 0.967 0.947 1 0.989 ...
 $ Outside_Global_Index : num  1 1 1 1 1 1 1 1 1 1 ...
 $ LogOfAreas           : num  2.43 2.03 1.85 2.25 3.38 ...
 $ Log_X_Index          : num  0.903 0.778 0.778 0.845 1.23 ...
 $ Log_Y_Index          : num  1.64 1.46 1.26 1.65 2.41 ...
 $ Orientation_Index    : num  0.818 0.793 0.667 0.844 0.934 ...
 $ Luminosity_Index     : num  -0.291 -0.176 -0.123 -0.157 -0.199 ...
 $ SigmoidOfAreas       : num  0.582 0.298 0.215 0.521 1 ...
 $ Pastry               : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Z_Scratch            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ K_Scatch             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Stains               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Dirtiness            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Bumps                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Other_Faults         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ type                 : Factor w/ 8 levels "0","Pastry","Z_Scratch",..: 2 2 2 2 2 2 2 2 2 2 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

運用此處參數建立模型，預測缺陷形態

要預測的變量

方法簡述

數據預整理

↓

數據分為訓練集和測試集

↓

以訓練集建立預測模型

↓

以測試集帶入模型測試準確性

↓

取得一模型，將參數輸入獲得預測值做參考

預測結果

簡單多變數回歸法，準確率=70%

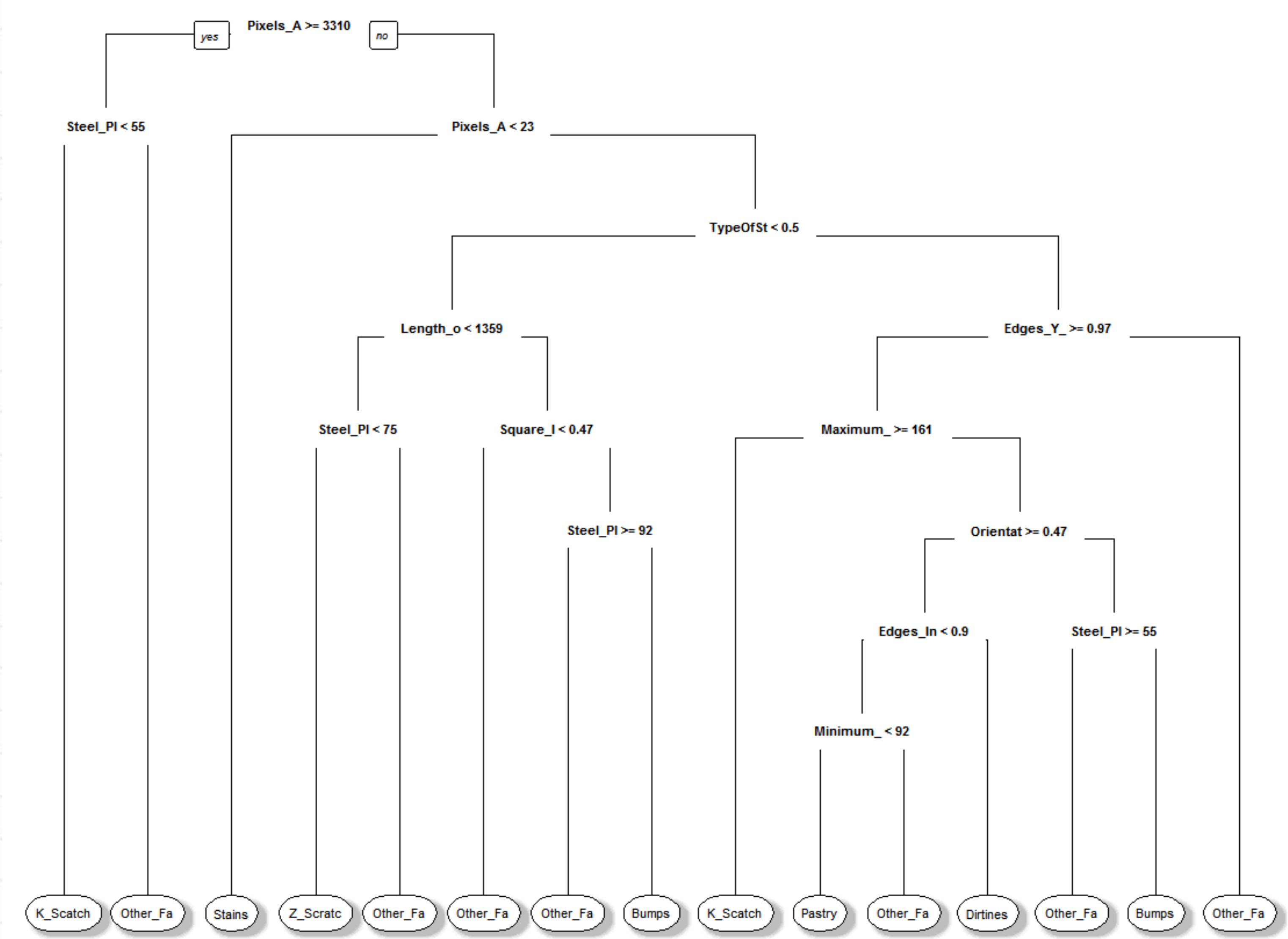
real \ predict	Pastry	Z_Scratch	K_Scatch	Stains	Dirtiness	Bumps	Other_Faults
Pastry	13	1	0	0	1	12	9
Z_Scratch	0	32	1	0	0	3	7
K_Scatch	0	0	75	1	0	2	3
Stains	0	0	0	14	0	1	0
Dirtiness	0	0	0	0	3	0	1
Bumps	2	2	0	0	0	52	21
Other_Faults	6	2	1	0	1	40	83

支援向量機法(SVM)，準確率=76%

real \ test.pred	Pastry	Z_Scratch	K_Scatch	Stains	Dirtiness	Bumps	Other_Faults
Pastry	18	0	0	0	0	6	12
Z_Scratch	0	34	0	0	0	2	7
K_Scatch	0	0	75	1	0	0	5
Stains	0	0	0	14	0	1	0
Dirtiness	0	0	0	0	4	0	0
Bumps	2	5	0	0	0	47	23
Other_Faults	3	1	1	1	0	23	104

預測結果

決策樹法



yes
Pixels_A >= 3310
no
Steel_Pl < 55
Pixels_A < 23
TypeOfSt < 0.5
Length_o < 1359
Edges_Y_ >= 0.97
Steel_Pl < 75
Square_I < 0.47
Maximum_ >= 161
Steel_Pl >= 92
Orientat >= 0.47
Edges_In < 0.9
Steel_Pl >= 55
Minimum_ < 92
K_Scatch
Other_Fa
Stains
Z_Scratc
Other_Fa
Other_Fa
Other_Fa
Bumps
K_Scatch
Pastry
Other_Fa
Dirtines
Other_Fa
Bumps
Other_Fa

預測結果

決策樹法，準確率=70%

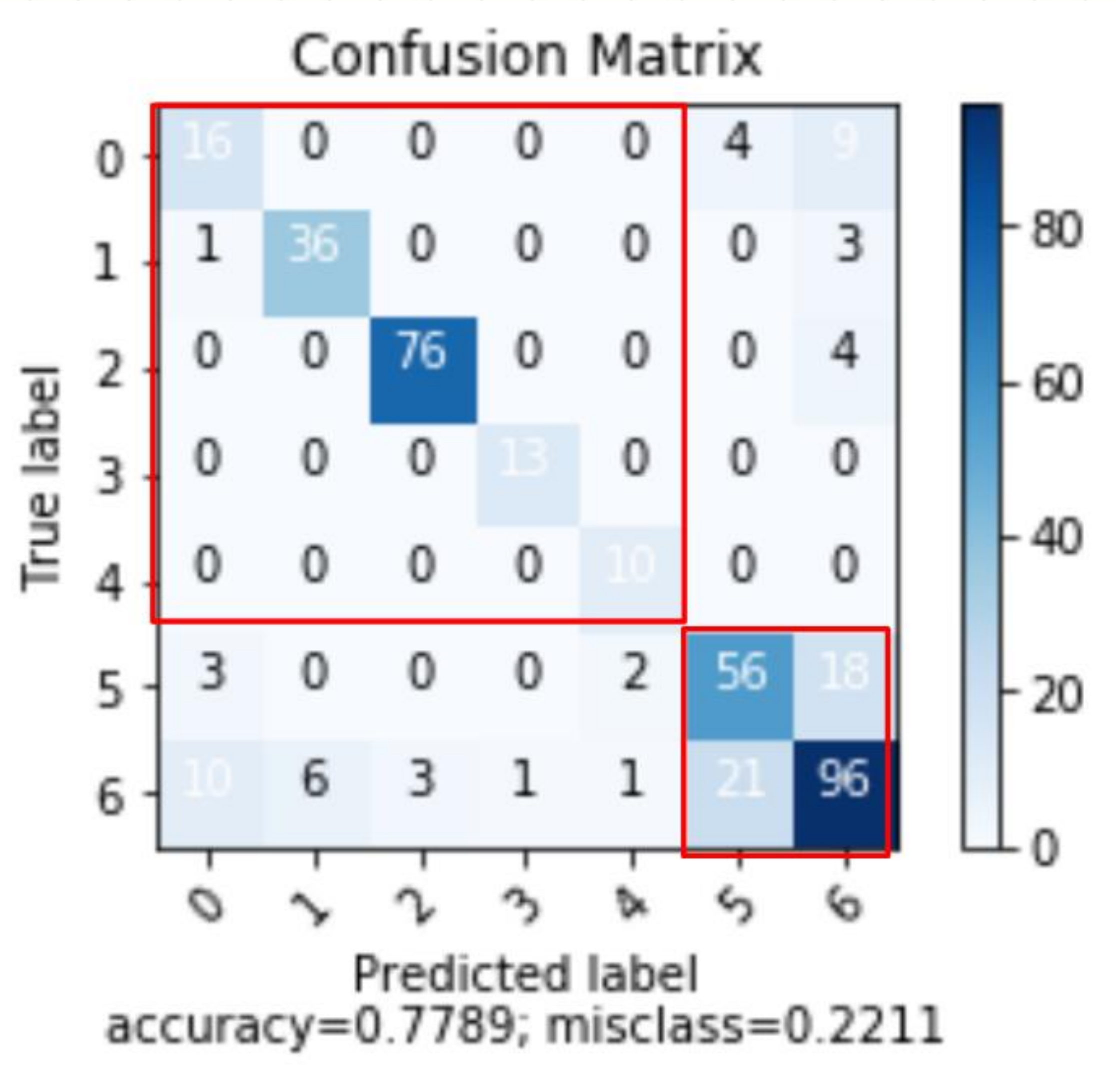
real \ predict	Pastry	Z_Scratch	K_Scatch	Stains	Dirtiness	Bumps	Other_Faults
Pastry	14	2	2	0	0	2	16
Z_Scratch	0	33	0	0	0	0	10
K_Scatch	0	0	70	1	0	2	8
Stains	0	0	0	12	0	2	1
Dirtiness	0	0	0	0	2	0	2
Bumps	1	1	0	0	0	46	29
Other_Faults	3	2	0	0	2	31	95

隨機森林法，準確率=80%

real \ predict	Pastry	Z_Scratch	K_Scatch	Stains	Dirtiness	Bumps	Other_Faults
Pastry	21	0	1	0	0	3	11
Z_Scratch	0	37	1	0	0	0	5
K_Scatch	0	0	79	0	0	0	2
Stains	0	0	0	14	0	1	0
Dirtiness	0	0	0	0	4	0	0
Bumps	3	0	1	0	1	49	23
Other_Faults	8	0	0	0	0	19	106

預測結果

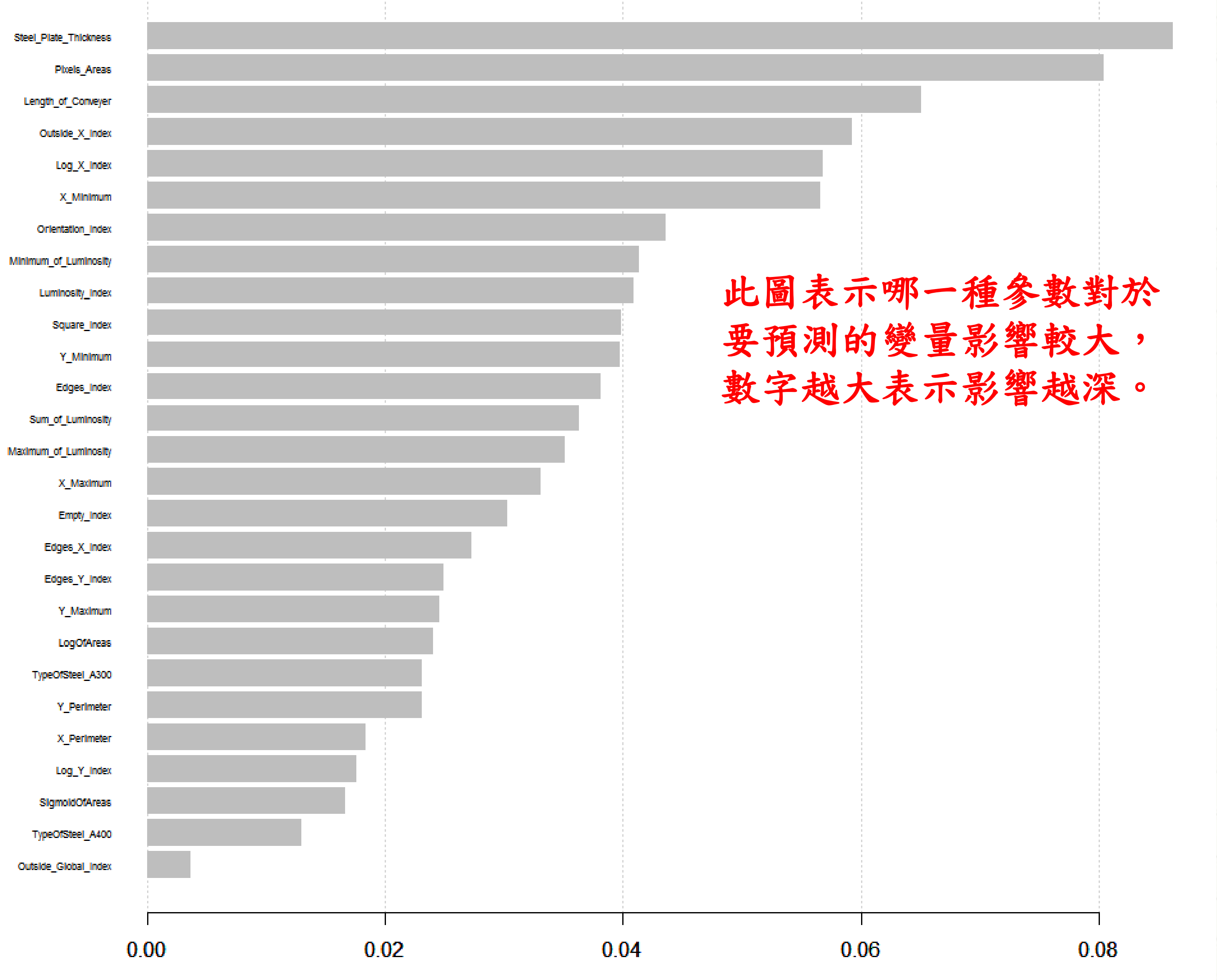
類神經網路法



此部分用Python撰寫，python在類神經網路語法齊全。

預測結果

xgboost



此圖表示哪一種參數對於要預測的變量影響較大，數字越大表示影響越深。

預測結果

xgboost，準確率=76　%

preidct \ real	Bumps	Dirtiness	K_Scatch	Other_Faults	Pastry	Stains	Z_Scratch
Bumps	64	0	0	31	8	2	0
Dirtiness	1	15	0	2	2	0	0
K_Scatch	0	0	76	2	0	1	0
Other_Faults	26	0	2	145	9	1	3
Pastry	5	0	0	17	19	0	1
Stains	0	0	0	1	0	11	0
Z_Scratch	0	0	1	1	0	0	40

第二部分：資料內容

```
Classes ‘data.table’ and 'data.frame':	4848 obs. of  29 variables:
 $ 鄉鎮市區                    : chr  "The villages and towns urban district" "中正區" "中正區" "中正區" ...
 $ 交易標的                    : chr  "transaction sign" "車位" "房地(土地+建物)" "房地(土地+建物)" ...
 $ 土地區段位置/建物區段門牌: chr  "land sector position/building sector house number plate" "臺北市中正區和平西路一段61~90號"
 $ 土地移轉總面積(平方公尺) : chr  "land shifting total area (square meter)" "0.28" "0.02" "6.72" ...
 $ 都市土地使用分區            : chr  "the use zoning or compiles and checks" "商" "住" "住" ...
 $ 非都市土地使用分區          : chr  "the non-metropolis land use district" NA NA NA ...
 $ 非都市土地使用編定          : chr  "non-metropolis land use" NA NA NA ...
 $ 交易年月日                  : chr  "transaction year" "1051024" "1051024" "1051007" ...
 $ 交易筆棟數                  : chr  "month and day" "土地0建物0車位1" "土地1建物1車位0" "土地3建物1車位0" ...
 $ 移轉層次                    : chr  "transaction pen number" "地下一層，地下二層，地下三層，地下四層" "五層" "五層" ...
 $ 總樓層數                    : chr  "shifting level" "二十八層" "五層" "十二層" ...
 $ 建物型態                    : chr  "total floor number" "其他" "華廈(10層含以下有電梯)" "住宅大樓(11層含以上有電梯)" ...
 $ 主要用途                    : chr  "building state" "停車空間" NA "住家用" ...
 $ 主要建材                    : chr  "main use" "鋼骨鋼筋混凝土造" "加強磚造" "鋼筋混凝土造" ...
 $ 建築完成年月                : chr  "main building materials" "930120" "560519" "681226" ...
 $ 建物移轉總面積(平方公尺) : chr  "construction to complete the years" "40.77" "42.72" "72.47" ...
 $ 建物現況格局-房             : chr  "building shifting total area" "0" "3" "2" ...
 $ 建物現況格局-廳             : chr  "Building present situation pattern - room" "0" "1" "2" ...
 $ 建物現況格局-衛             : chr  "building present situation pattern - hall" "0" "1" "1" ...
 $ 建物現況格局-隔間           : chr  "building present situation pattern - health" "有" "有" "有" ...
 $ 有無管理組織                : chr  "building present situation pattern - compartmented" "無" "有" "有" ...
 $ 總價(元)                    : chr  "whether there is manages the organization" "3100000" "51983" "12250000" ...
 $ 單價(元/平方公尺)           : chr  "total price (Yuan)" NA "1217" "169035" ...
 $ 車位類別                    : chr  "the unit price (a Yuan/square meter)" "坡道平面" NA NA ...
 $ 車位移轉總面積(平方公尺) : chr  "the berth category" "0" "0" "0" ...
 $ 車位總價(元)                : chr  "berth shifting total area (square meter)" "3100000" "0" "0" ...
 $ 備註                        : chr  "the berth total price (Yuan)" NA "親友、員工或其他特殊關係間之交易。" NA ...
 $ 編號                        : chr  "the note" "RPSNMLMJJILFFAA96CA" "RPSNMLNJJILFFAA07CA" "RPPNMLPJJILFFAA96CA" ...
 $ V29                         : chr  "serial number" NA NA NA ...
```

挑選適當參數以建立預測總價的模型

方法簡述

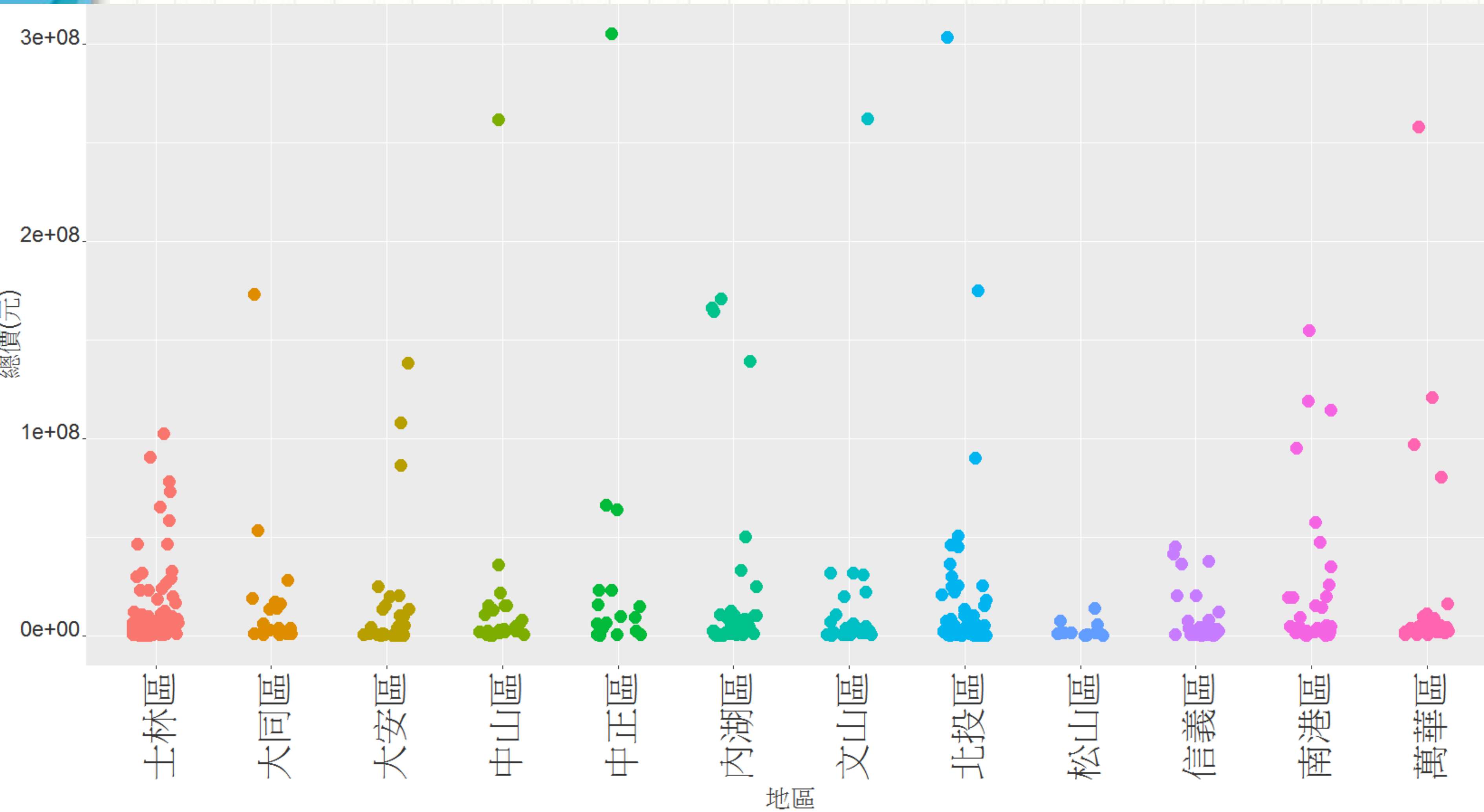
數據預分類 → 數據預處理 → 數據分為訓練集和測試集

依照交易標的先分類
1. 土地
2. 車位
3. 土地+建物
4. 土地+建物+車位
5. 建物

1. 處理遺漏值
2. 處理離群值
3. 選擇適當變數

以訓練集建立預測模型 → 以測試集帶入模型測試準確性 → 取得一模型，將參數輸入獲得預測值做參考

資料視覺化

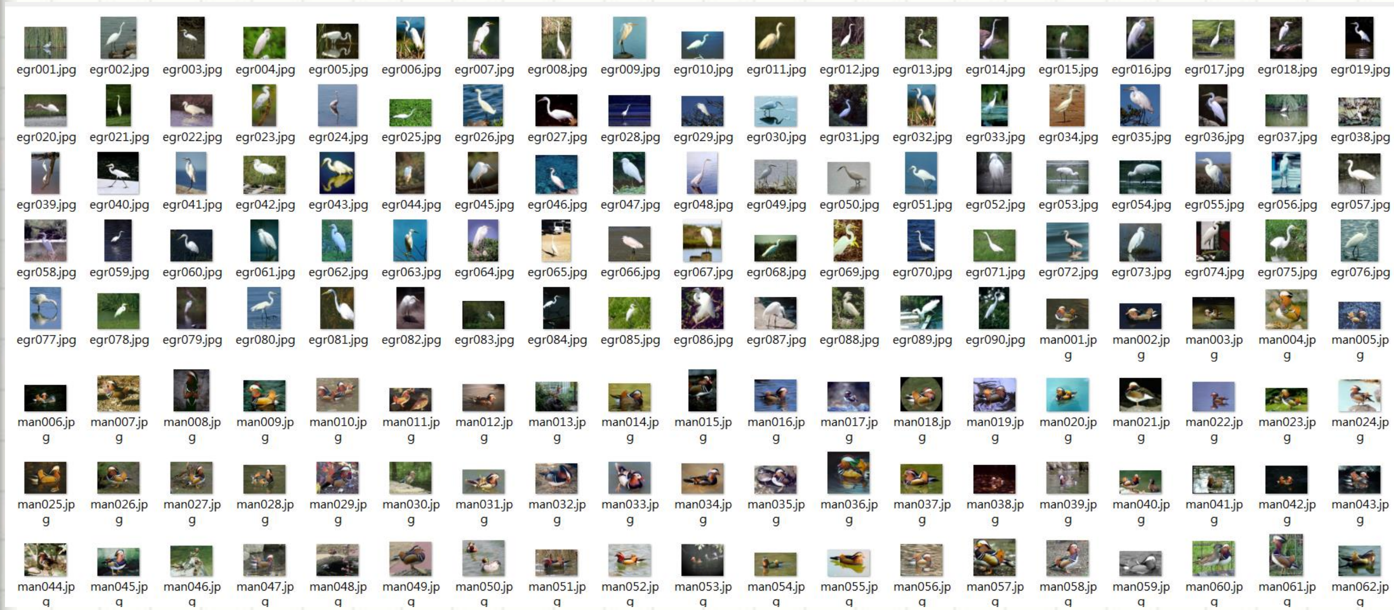


本圖表說明不同地區其不同交易價格的分布

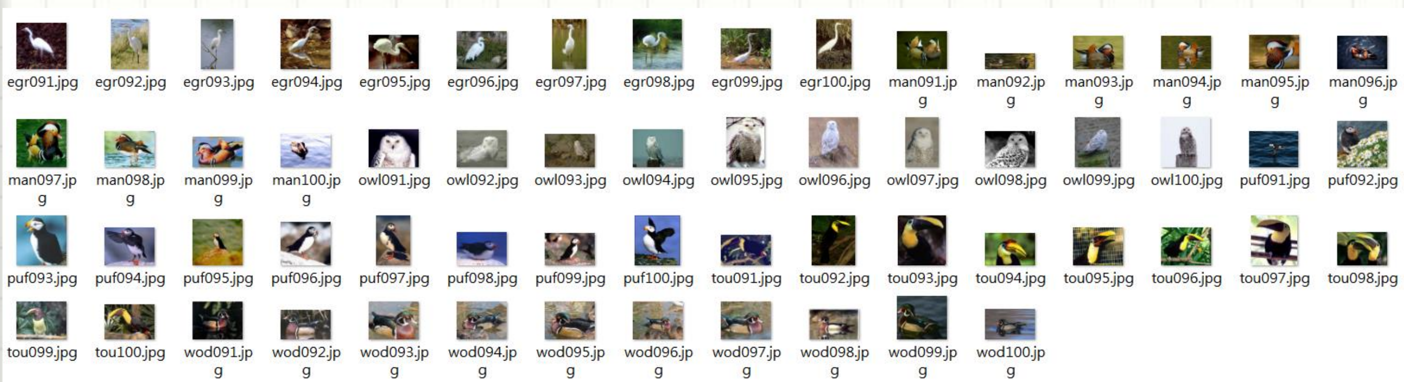
預測結果

方法	預測項目	均方根誤差
Lasso回歸	土地	15,553,052
xgboost	土地	3,427,291
支援向量機法(SVM)	土地	7,985,620

第三部分：資料內容(1/2)



訓練集圖檔



測試集圖檔

第三部分：資料內容

利用CNN(卷積神經網路Convolutional Neural Networks)訓練模型做不同種類鳥之辨識。






方法簡述

移除不適當圖，
分為訓練集和測試集

→

以Keras建立
CNN模型

1. 擷取各圖的特徵做為變數
2. 建立適當類神經網路
3. 選擇適當參數(激發函數、優化器)

以測試集帶入
模型測試準確性

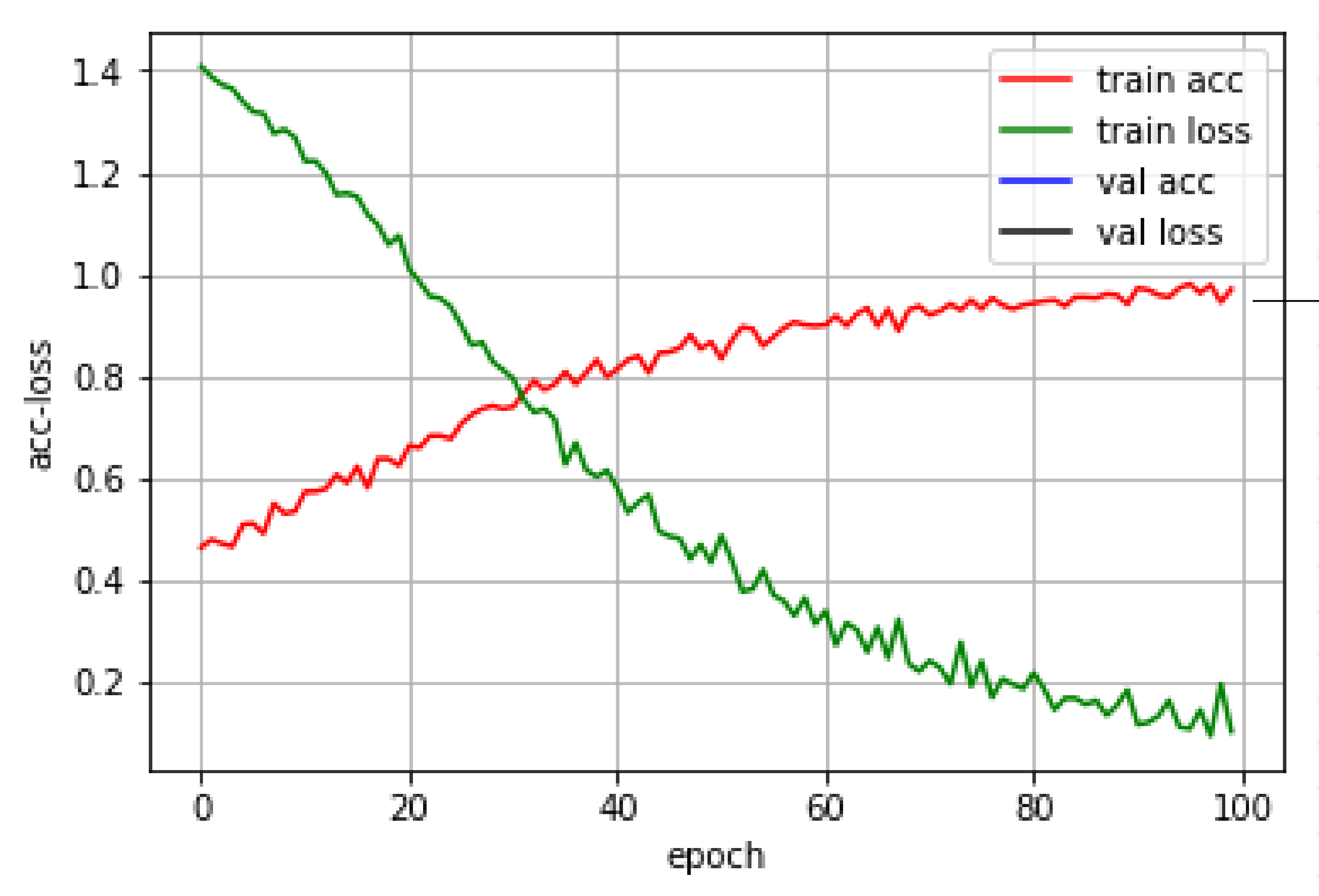
→

取得一模型，將
參數輸入獲得預
測之類別做參考

預測結果

模型A

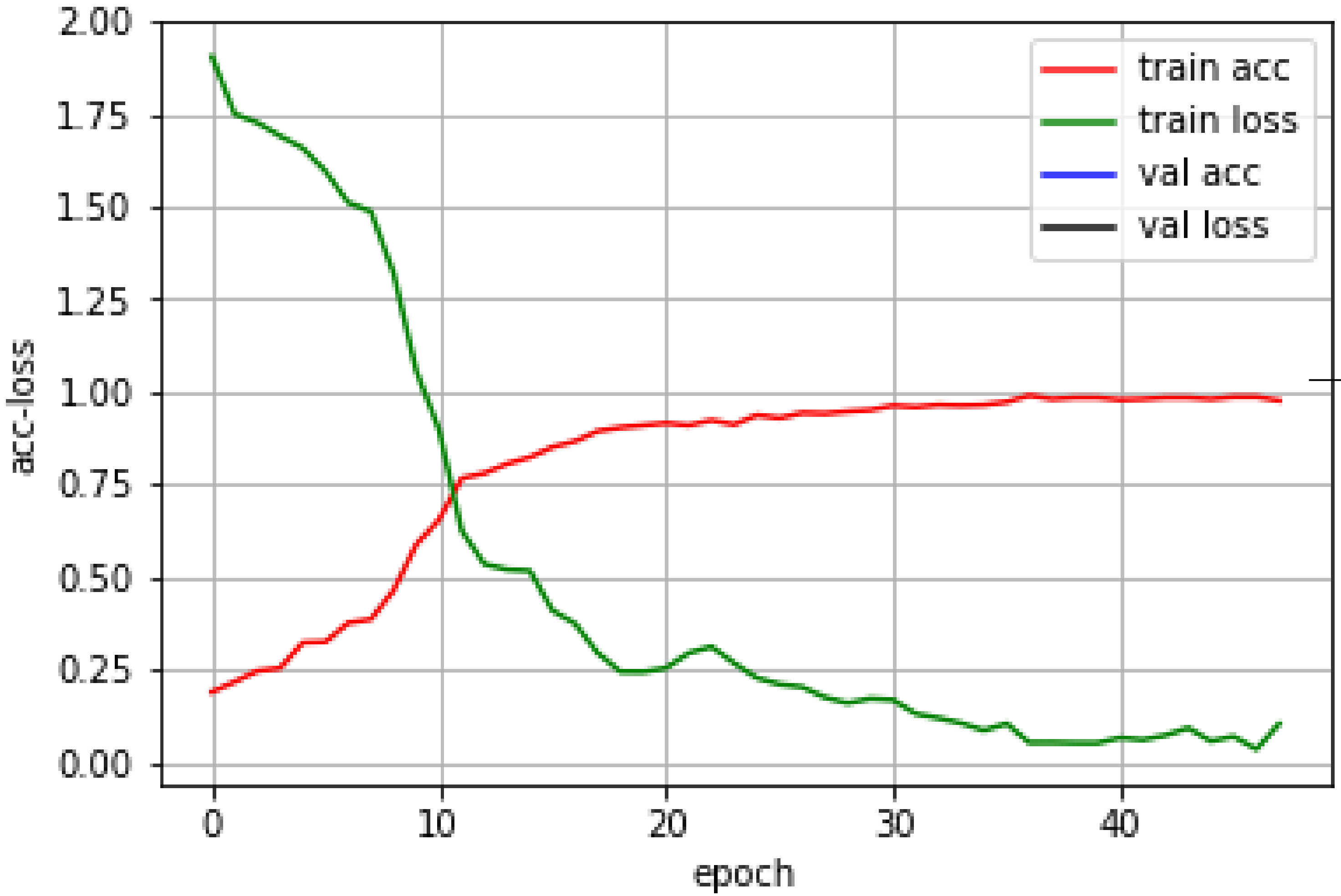
訓練越多次準確率
越高(針對訓練集圖)



預測結果

模型B

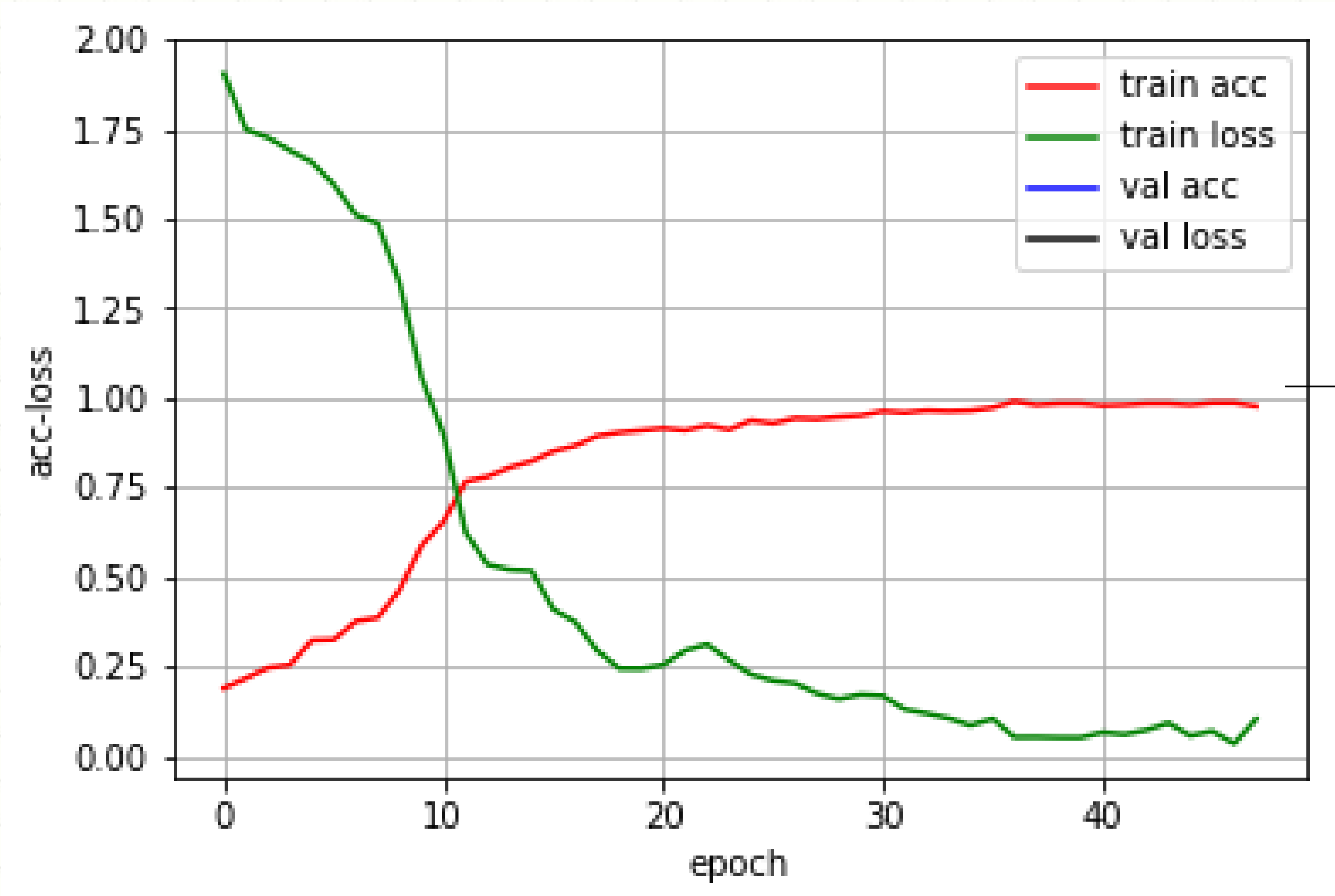
用其他優化器，在運行50次後得到之結果



預測結果

模型B

用其他優化器，在運行50次後得到之結果



預測結果

將測試的圖片以
模型B去判讀
對角線為正確預
測的部分
正確率：37%

predict	0	1	2	3	4	5
label						
0.0	3	0	3	0	4	2
1.0	2	5	3	2	1	0
2.0	0	1	3	3	0	4
3.0	3	2	2	4	0	0
4.0	2	0	1	0	8	1
5.0	0	0	4	1	3	3

預測結果

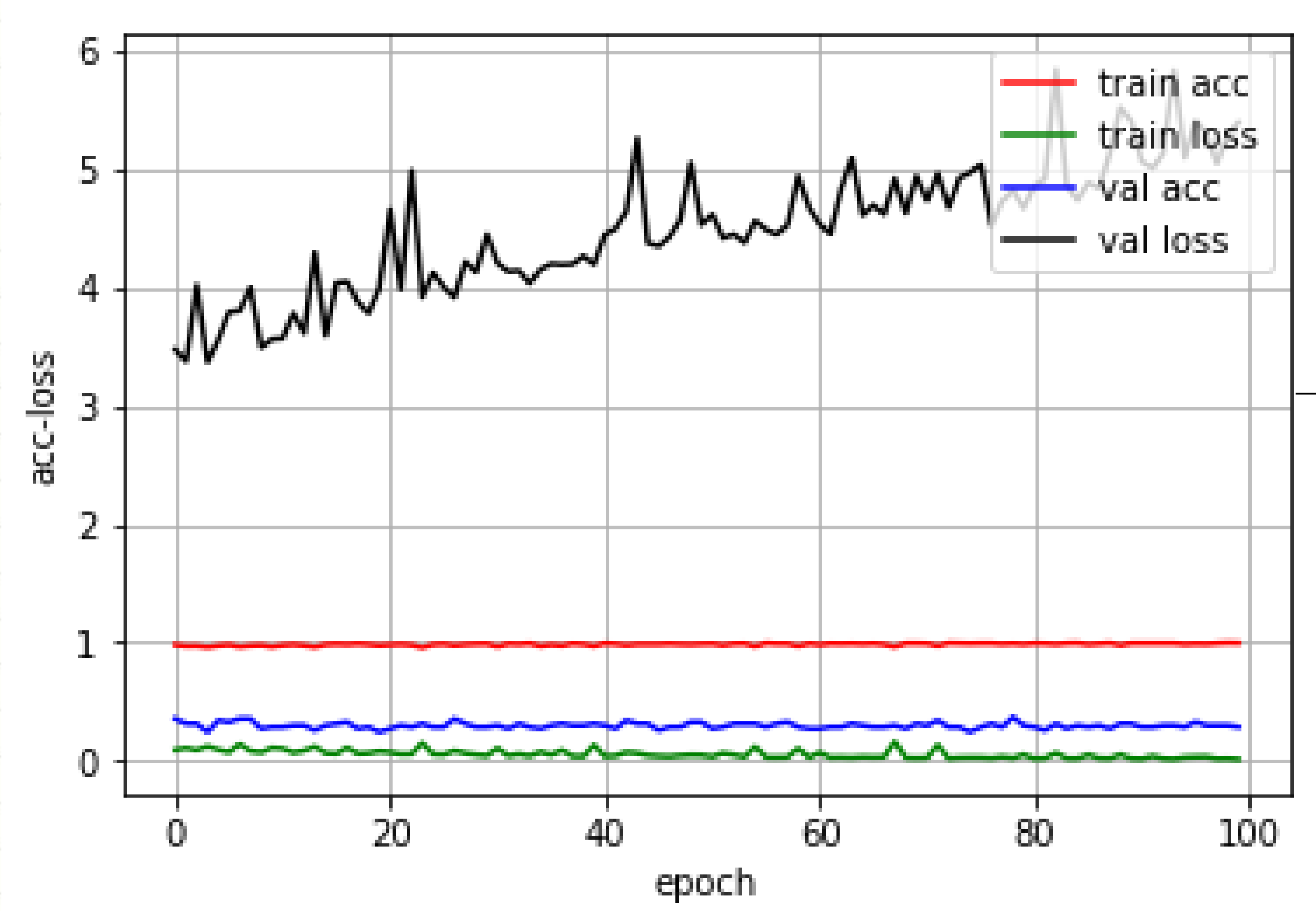
亦可以從其他預測錯誤的案例來推知，當模型預測為2時，較不容易辨別實際為哪一類型。

predict label	0	1	2	3	4	5
0.0	3	0	3	0	4	2
1.0	2	5	3	2	1	0
2.0	0	1	3	3	0	4
3.0	3	2	2	4	0	0
4.0	2	0	1	0	8	1
5.0	0	0	4	1	3	3

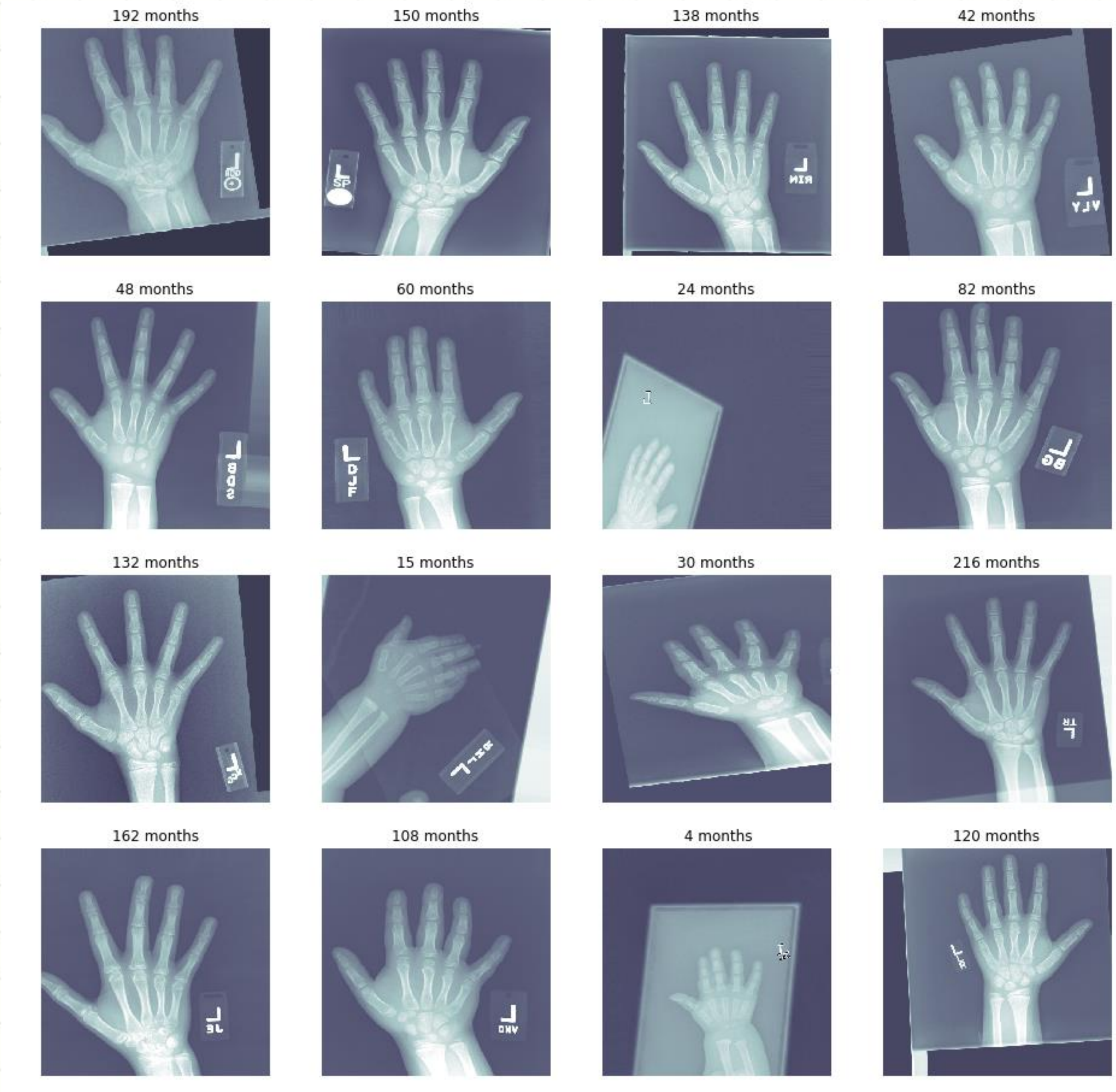
預測結果

模型C

用其他優化器，在運行50次後得到之結果



第三部分：資料內容(2/2)



目標：學習圖片中特徵預測該手的年齡。

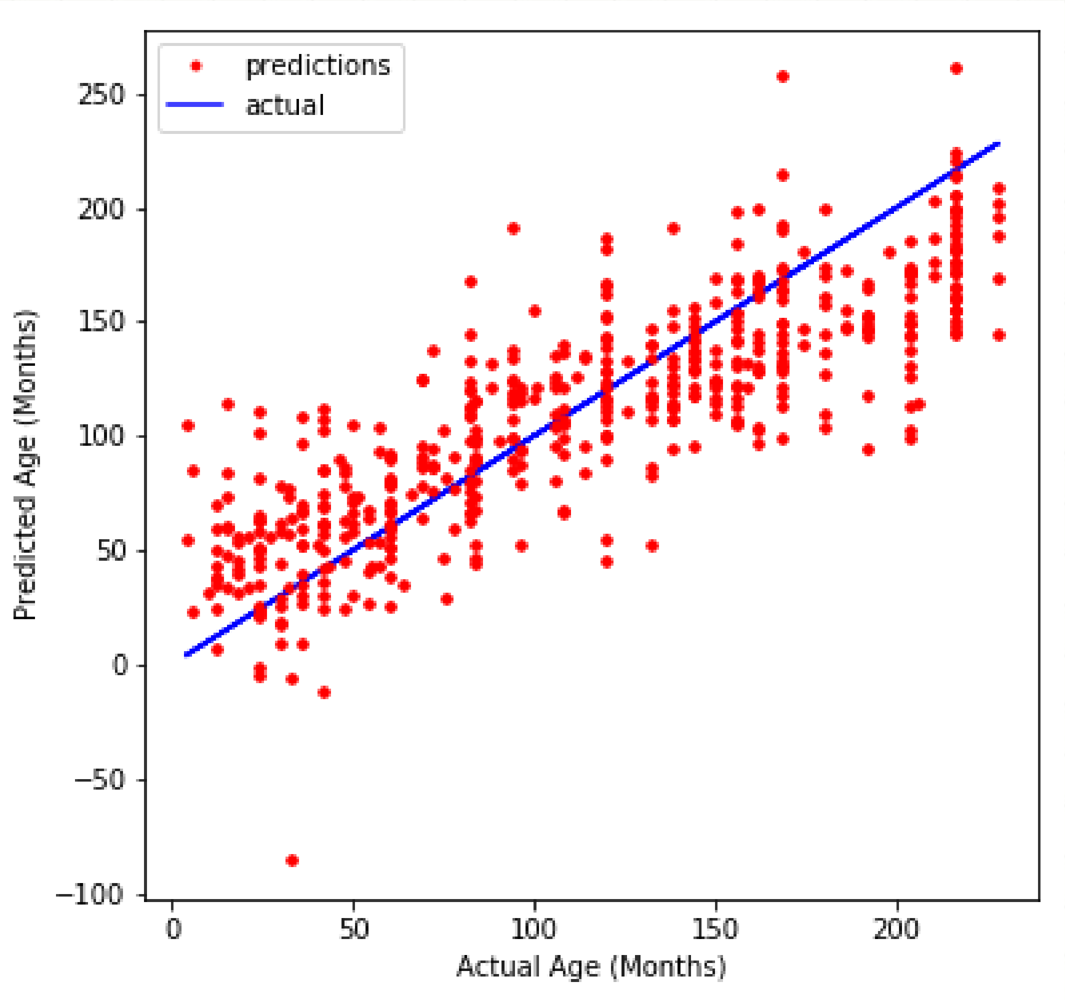
預測結果

運算過程：運算兩周期，共約22小時。

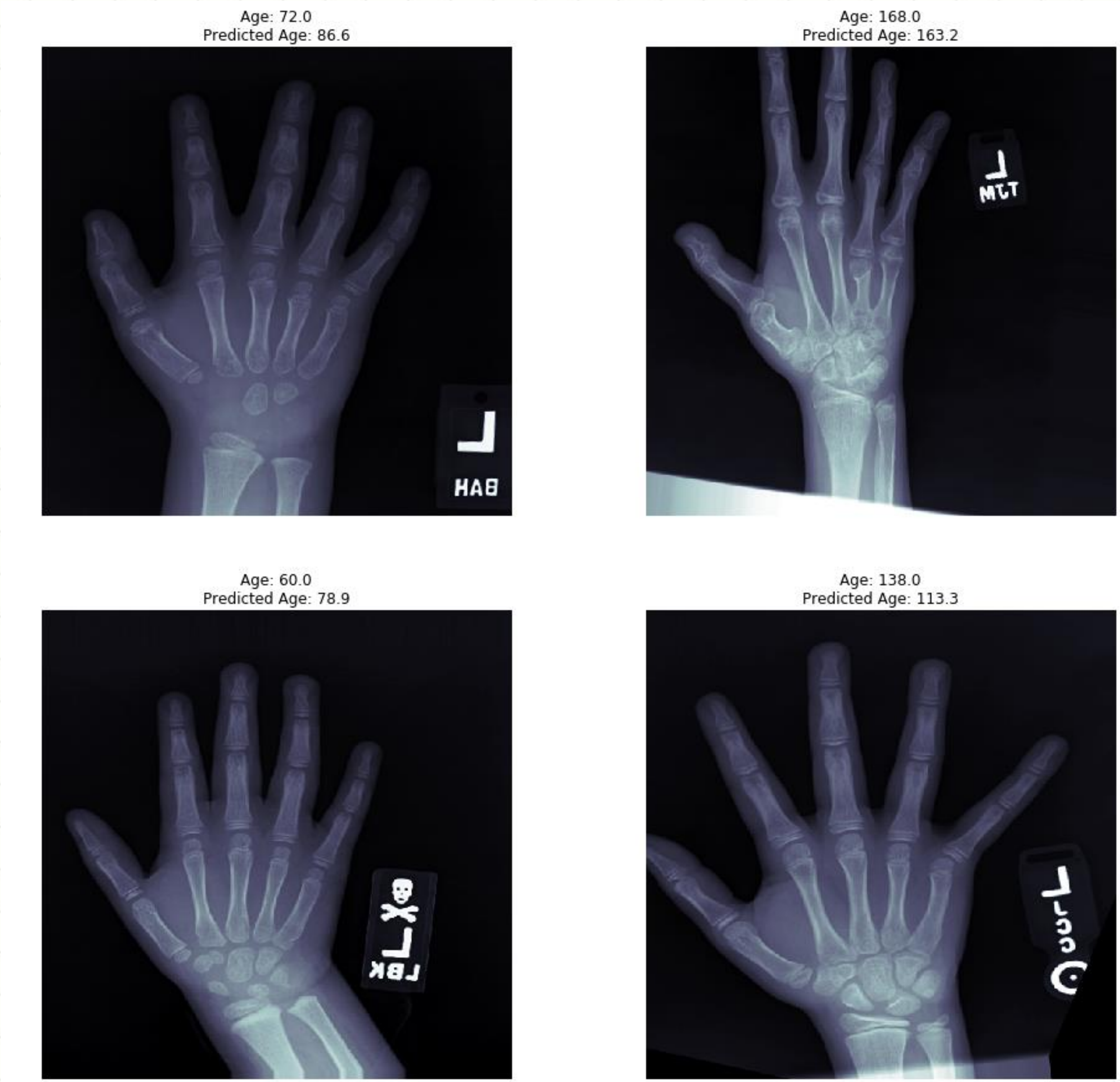
```
Epoch 1/2
469/469 [==============================] - 56076s 120s/step - loss: 0.4437 - mae_months: 41.8415 - val_loss: 2.7369 - val_mae_months: 119.4757

Epoch 00001: val_loss improved from inf to 2.73690, saving model to bone_age_weights.best.hdf5
Epoch 2/2
469/469 [==============================] - 26490s 56s/step - loss: 0.2009 - mae_months: 28.1547 - val_loss: 0.1959 - val_mae_months: 27.7873
```

預測結果

預測值對實際值作圖



預測結果


Age: 72.0
Predicted Age: 86.6
Age: 168.0
Predicted Age: 163.2
Age: 60.0
Predicted Age: 78.9
Age: 138.0
Predicted Age: 113.3

預測結果

Age: 216.0
Predicted Age: 199.9

Age: 42.0
Predicted Age: 61.9

Age: 50.0
Predicted Age: 61.8

Age: 36.0
Predicted Age: 97.3

預測結果

運算過程：運算三周期，共約44小時。

Epoch 1/3
469/469 [==============================] - 52020s 111s/step - loss: 0.3958 - mae_months: 38.7718 - val_loss: 0.8370 - val_mae_months: 64.4920

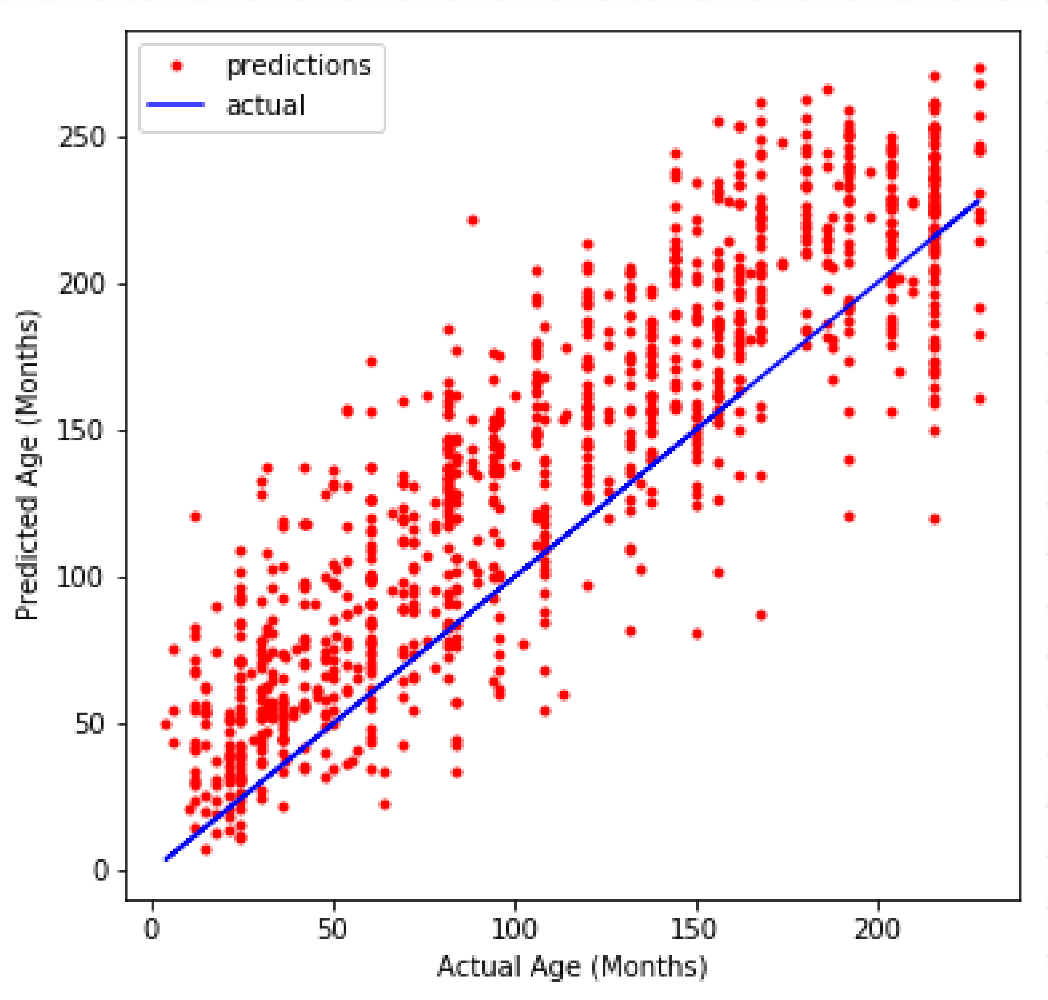
Epoch 00001: val_loss improved from inf to 0.83697, saving model to bone_age_weights.best.hdf5
Epoch 2/3
469/469 [==============================] - 55172s 98s/step - loss: 0.1963 - mae_months: 27.8057 - val_loss: 0.2818 - val_mae_months: 34.6985

Epoch 00002: val_loss improved from 0.83697 to 0.28182, saving model to bone_age_weights.best.hdf5
Epoch 3/3
469/469 [==============================] - 52356s 112s/step - loss: 0.1473 - mae_months: 24.1946 - val_loss: 0.3845 - val_mae_months: 42.2493

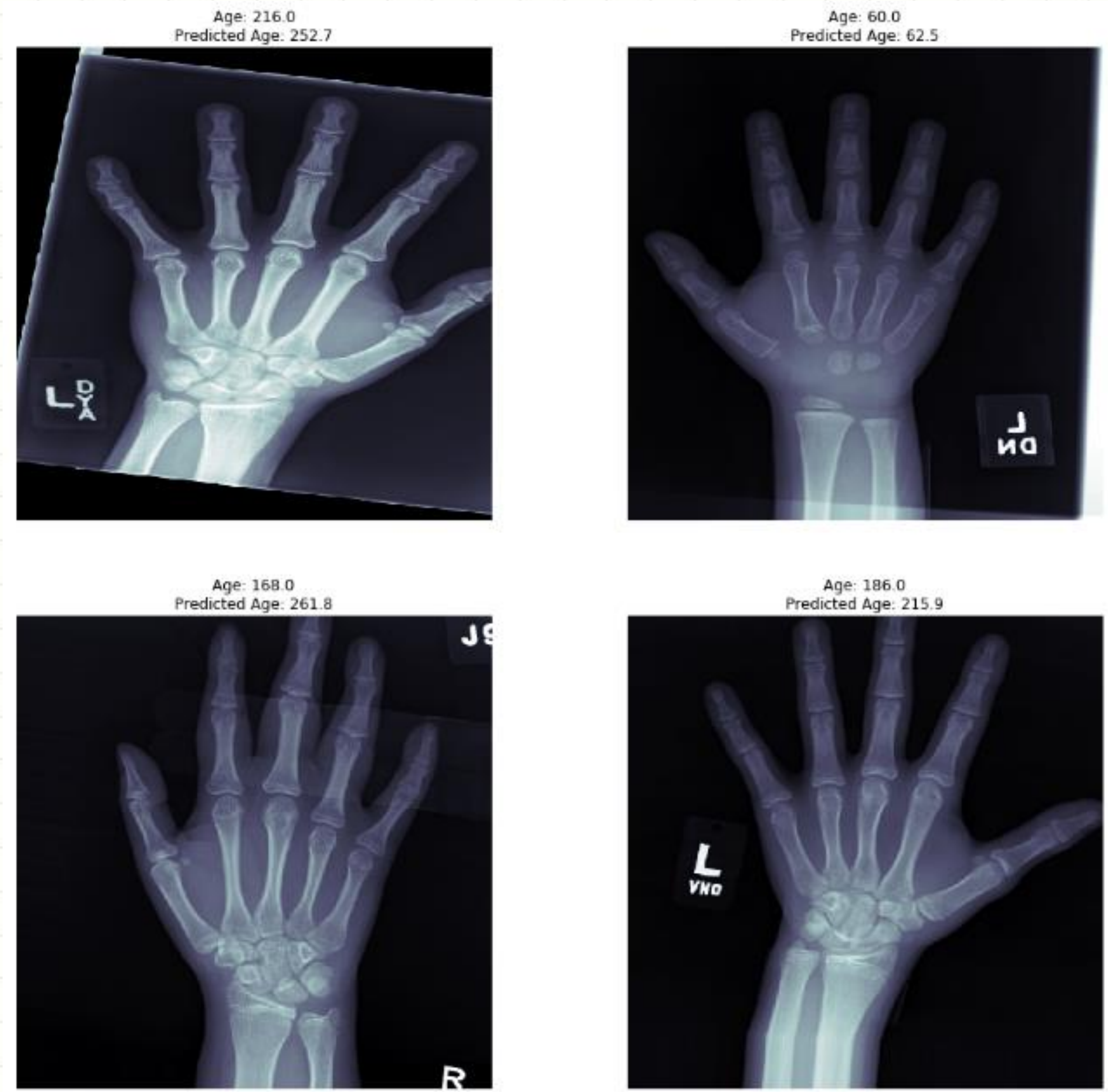
由以上可以看出模型訓練準確度提高，但是預測未知圖片時誤差變大，表示已經過度學習，若要再精準預測則要考慮調整其他參數，而非增加訓練時間。

預測結果

預測值對實際值作圖:與26頁圖相比誤差較大



預測結果



Age: 216.0
Predicted Age: 252.7
Age: 60.0
Predicted Age: 62.5
Age: 168.0
Predicted Age: 261.8
Age: 186.0
Predicted Age: 215.9

預測結果


Age: 60.0
Predicted Age: 85.1
Age: 60.0
Predicted Age: 100.0
Age: 54.0
Predicted Age: 36.3
Age: 150.0
Predicted Age: 143.2

預測結果

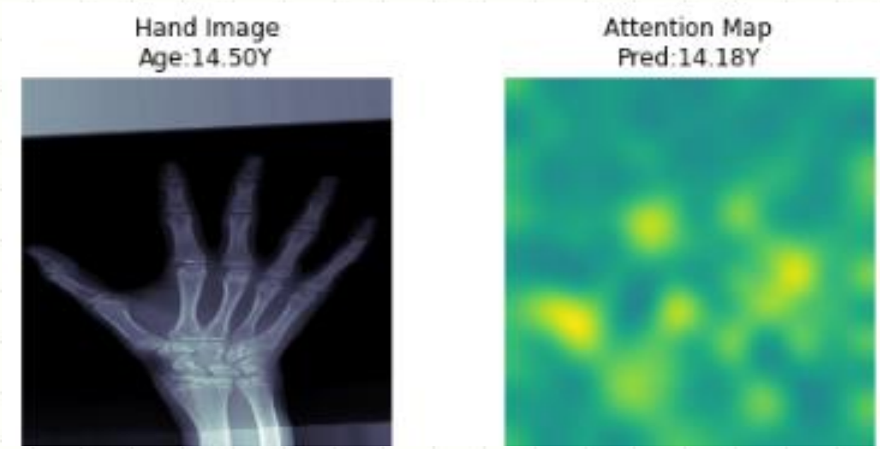
運算過程：運算六周期，共約12小時，採用另一種模型計算。

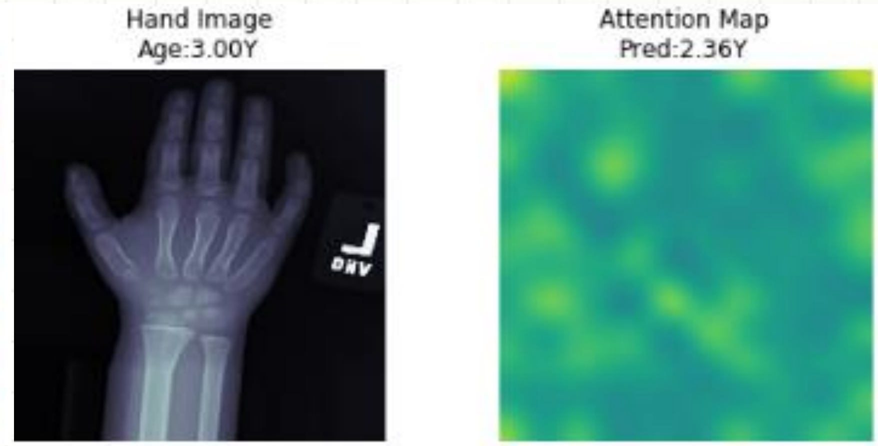
Epoch 1/6
235/235 [==============================] - 7525s 32s/step - loss: 0.1100 - mae_months: 21.0170 - val_loss: 0.0693 - val_mae_months: 16.8291
Epoch 00001: val_loss improved from 0.07636 to 0.06927, saving model to bone_age_weights.best.hdf5
Epoch 2/6
235/235 [==============================] - 7396s 31s/step - loss: 0.0947 - mae_months: 19.5161 - val_loss: 0.0692 - val_mae_months: 16.4281
Epoch 00002: val_loss improved from 0.06927 to 0.06918, saving model to bone_age_weights.best.hdf5
Epoch 3/6
235/235 [==============================] - 7435s 32s/step - loss: 0.0791 - mae_months: 17.7435 - val_loss: 0.0486 - val_mae_months: 13.9336
Epoch 00003: val_loss improved from 0.06918 to 0.04864, saving model to bone_age_weights.best.hdf5
Epoch 4/6
235/235 [==============================] - 7487s 32s/step - loss: 0.0720 - mae_months: 16.9528 - val_loss: 0.0480 - val_mae_months: 13.7648
Epoch 00004: val_loss improved from 0.04864 to 0.04800, saving model to bone_age_weights.best.hdf5
Epoch 5/6
235/235 [==============================] - 7589s 32s/step - loss: 0.0691 - mae_months: 16.6434 - val_loss: 0.0452 - val_mae_months:
13.2270

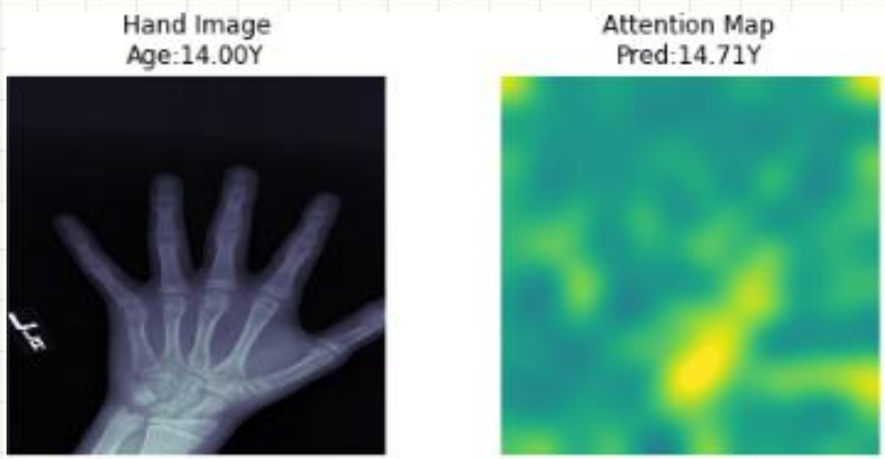
Epoch 00005: val_loss improved from 0.04800 to 0.04522, saving model to bone_age_weights.best.hdf5（較佳）
Epoch 6/6
235/235 [==============================] - 7627s 32s/step - loss: 0.0631 - mae_months: 15.7876 - val_loss: 0.0529 - val_mae_months: 14.2945

預測結果

將圖片之特徵繪製成attention map



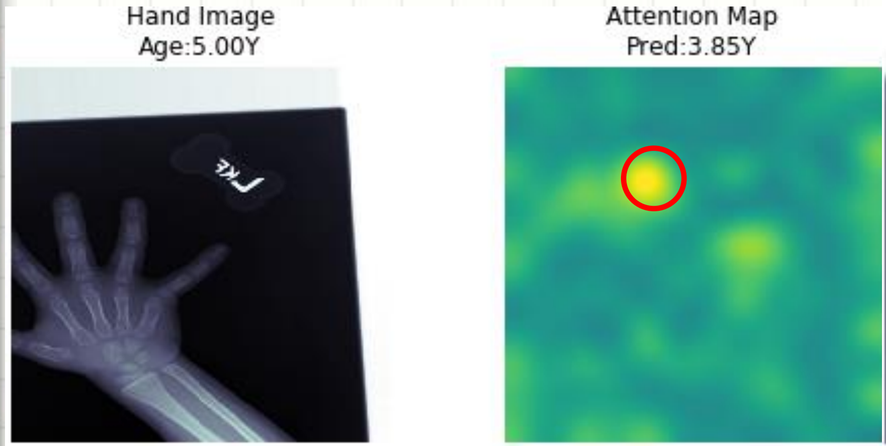


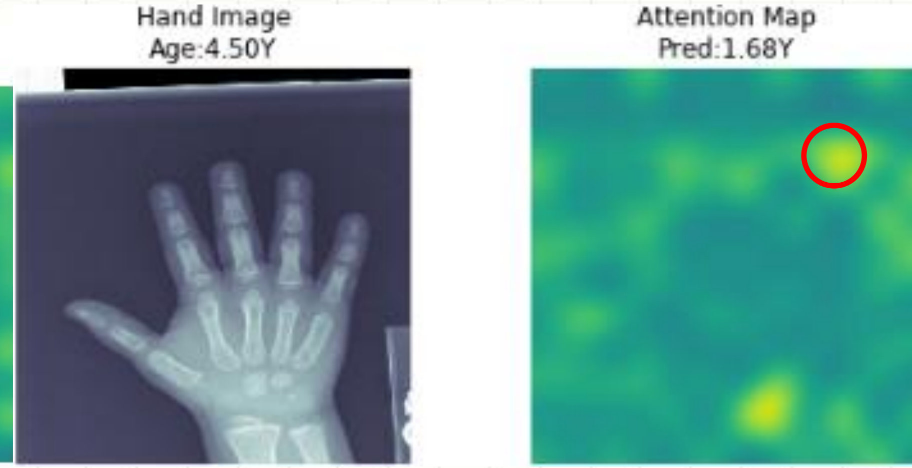


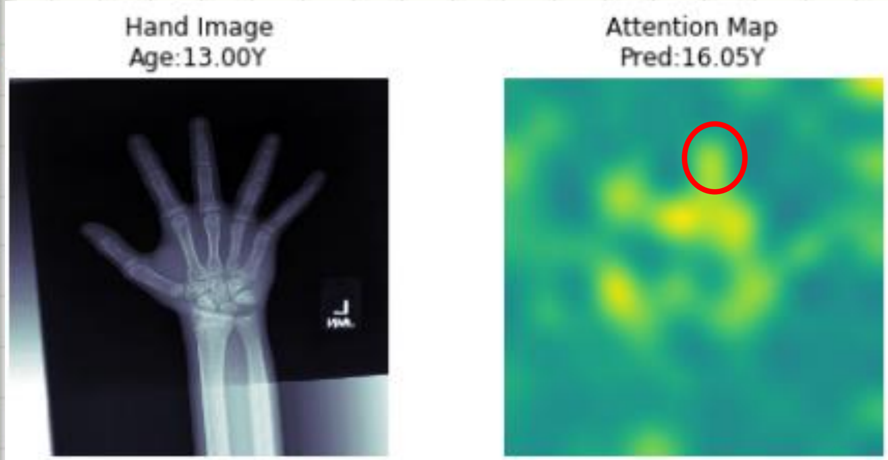
程式汲取的特徵以直觀來看與原圖較相符，這些案例所預測的值較準

預測結果

將圖片之特徵繪製成attention map



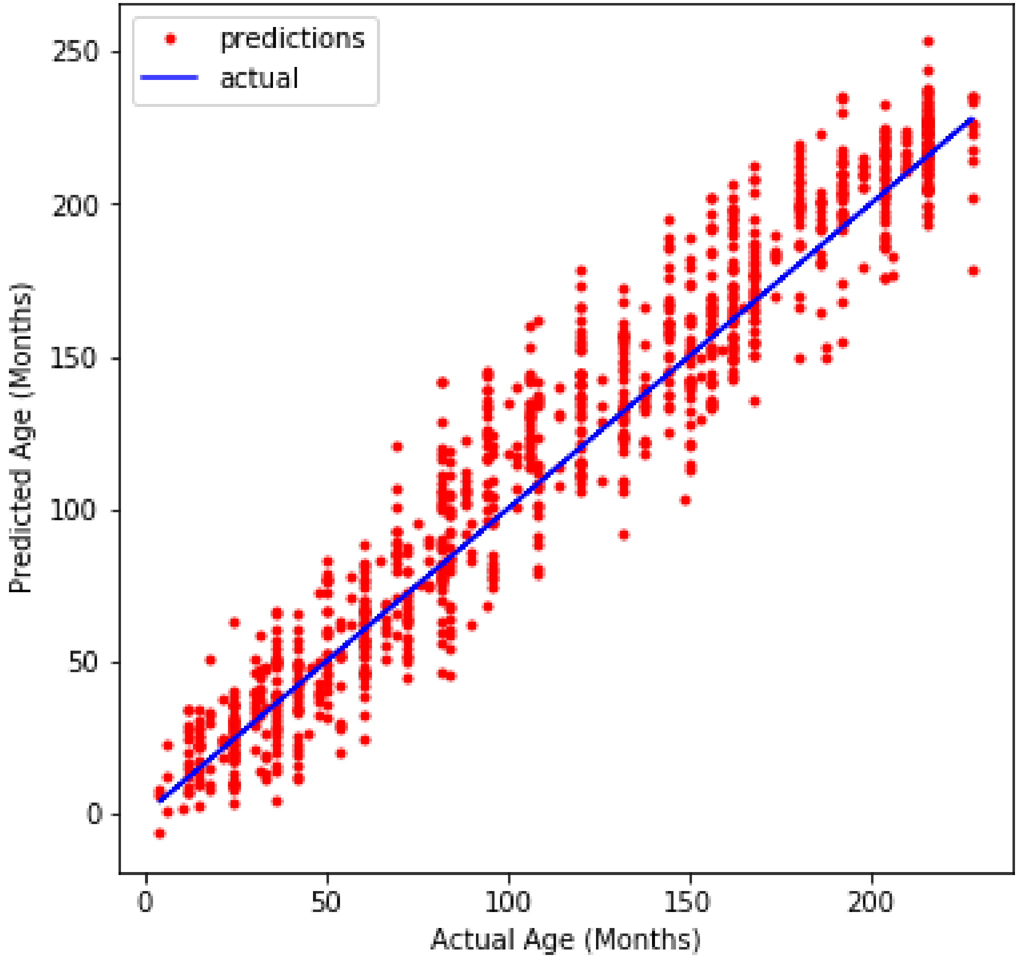




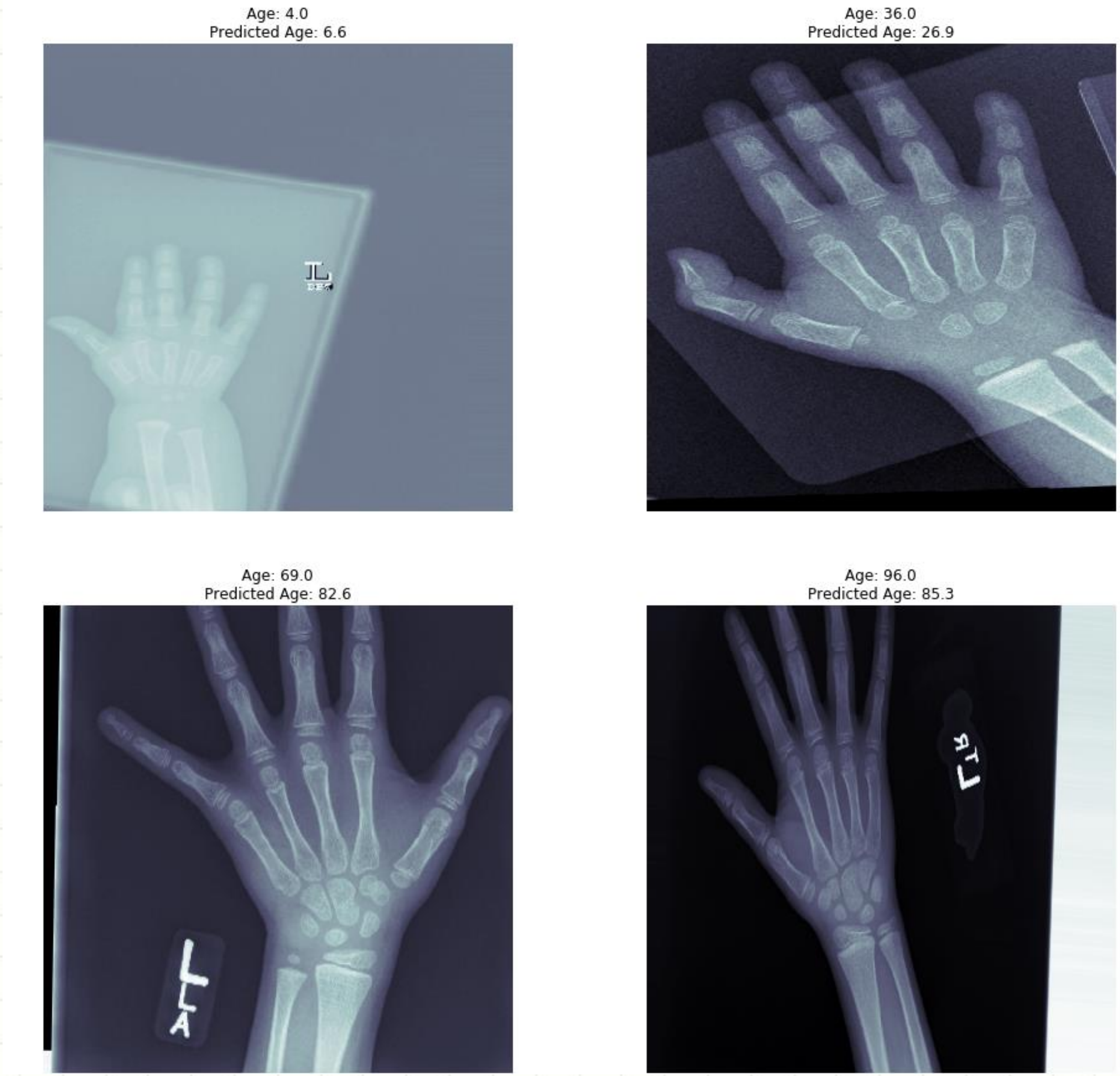
取得之特徵與原圖較比較無相關，
以這樣的特徵去預測所得到之值誤差較大

預測結果

預測值對實際值作圖:與其他圖相比誤差較小，MAE為13.2



預測結果


Age: 4.0
Predicted Age: 6.6
Age: 36.0
Predicted Age: 26.9
Age: 69.0
Predicted Age: 82.6
Age: 96.0
Predicted Age: 85.3

預測結果

Age: 132.0
Predicted Age: 119.4

Age: 162.0
Predicted Age: 154.3

Age: 198.0
Predicted Age: 212.6

Age: 228.0
Predicted Age: 214.3

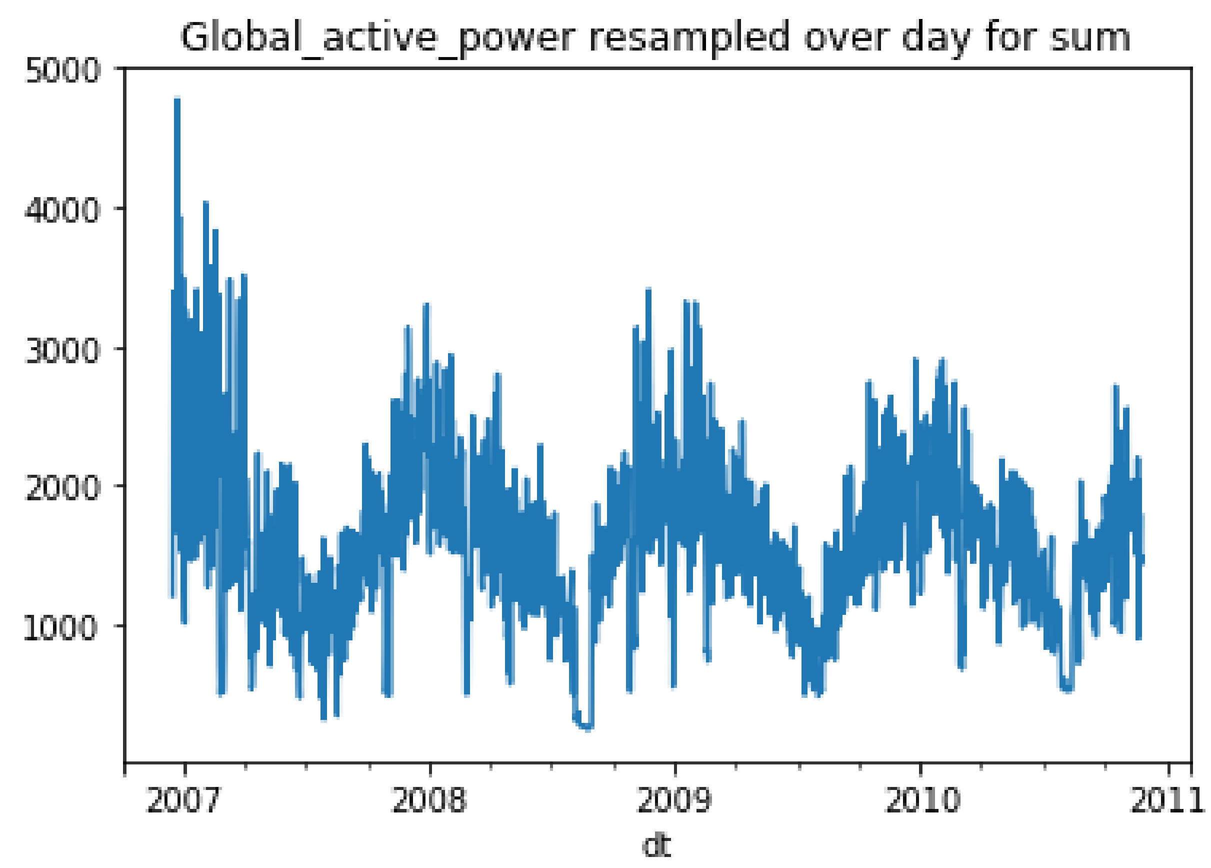
第四部分：資料內容

Index	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
2006-12-16 17:42:00	3.266	0	237.13	13.8	0	0	18
2006-12-16 17:43:00	3.728	0	235.84	16.4	0	0	17
2006-12-16 17:44:00	5.894	0	232.69	25.4	0	0	16
2006-12-16 17:45:00	7.706	0	230.98	33.2	0	0	17
2006-12-16 17:46:00	7.026	0	232.21	30.6	0	0	16
2006-12-16 17:47:00	5.174	0	234.19	22	0	0	17
2006-12-16 17:48:00	4.474	0	234.96	19.4	0	0	17
2006-12-16 17:49:00	3.248	0	236.66	13.6	0	0	17
2006-12-16 17:50:00	3.236	0	235.84	13.6	0	0	17
2006-12-16 17:51:00	3.228	0	235.6	13.6	0	0	17
2006-12-16 17:52:00	3.258	0	235.49	13.8	0	0	17

目標：以LSTM法，透過學習其他參數的特徵搭配時間序列，預測Global_active_power在不同時間的值。

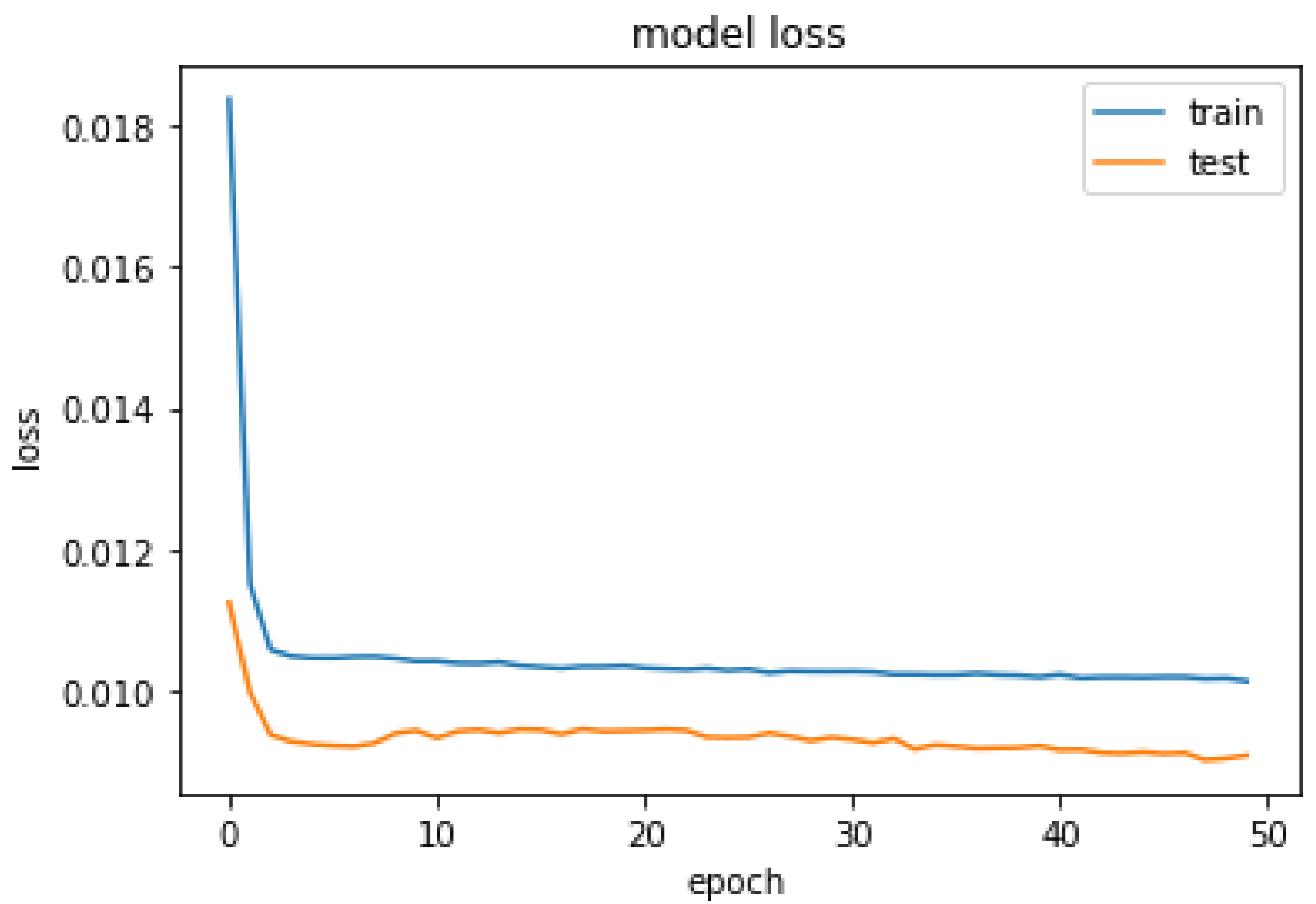
此範例在營業額對時間(月、季…)或是某原物料價隨時間變化上應用廣泛。

資料內容



Global_active_power對時間做圖，此參數是要預測的對象

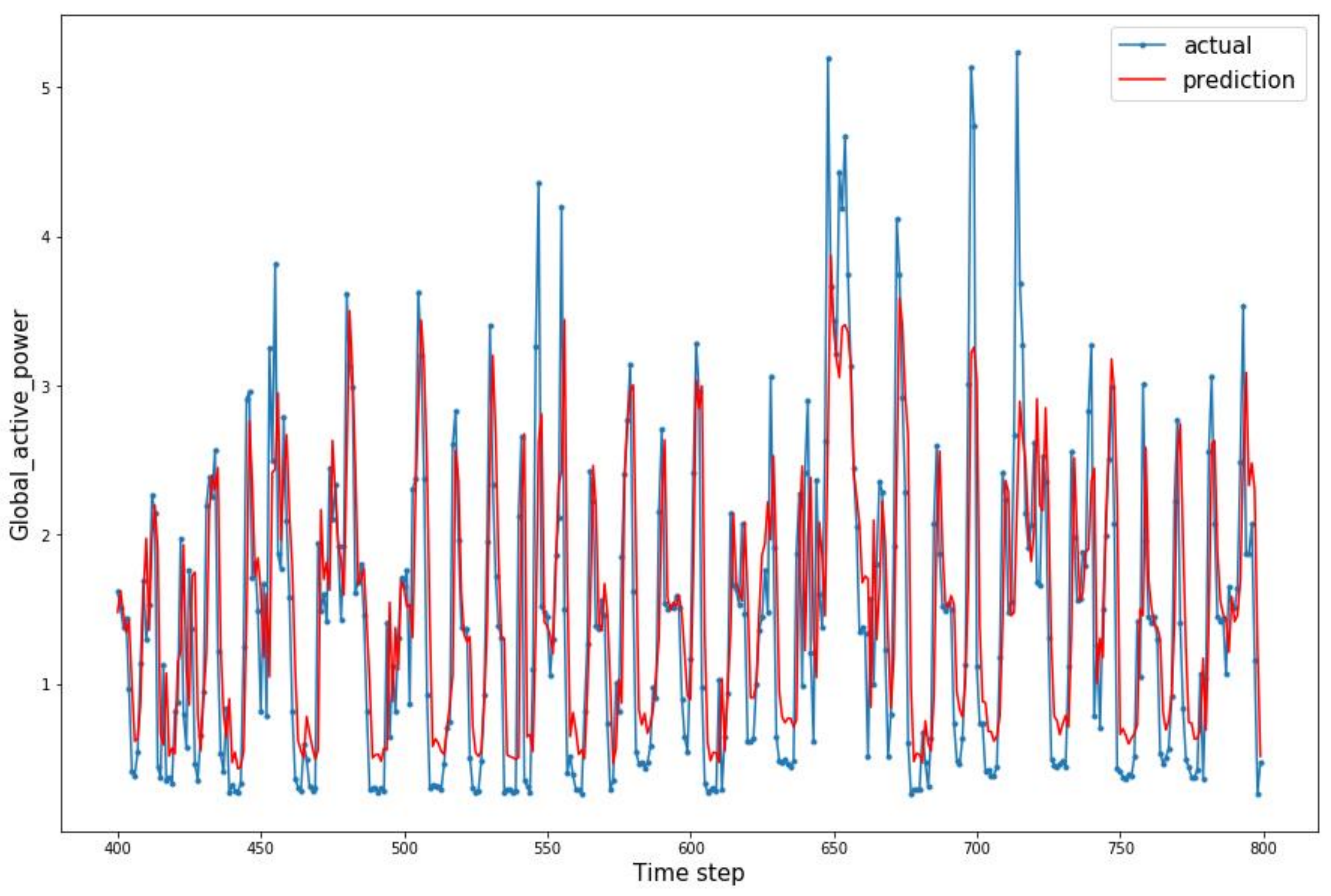
預測結果



預測結果

```
In [102]: print('Test RMSE: %.3f' % rmse)
Test RMSE: 0.614
```

預測結果


actual
prediction
Global_active_power
Time step
1
2
3
4
5
400
450
500
550
600
650
700
750
800

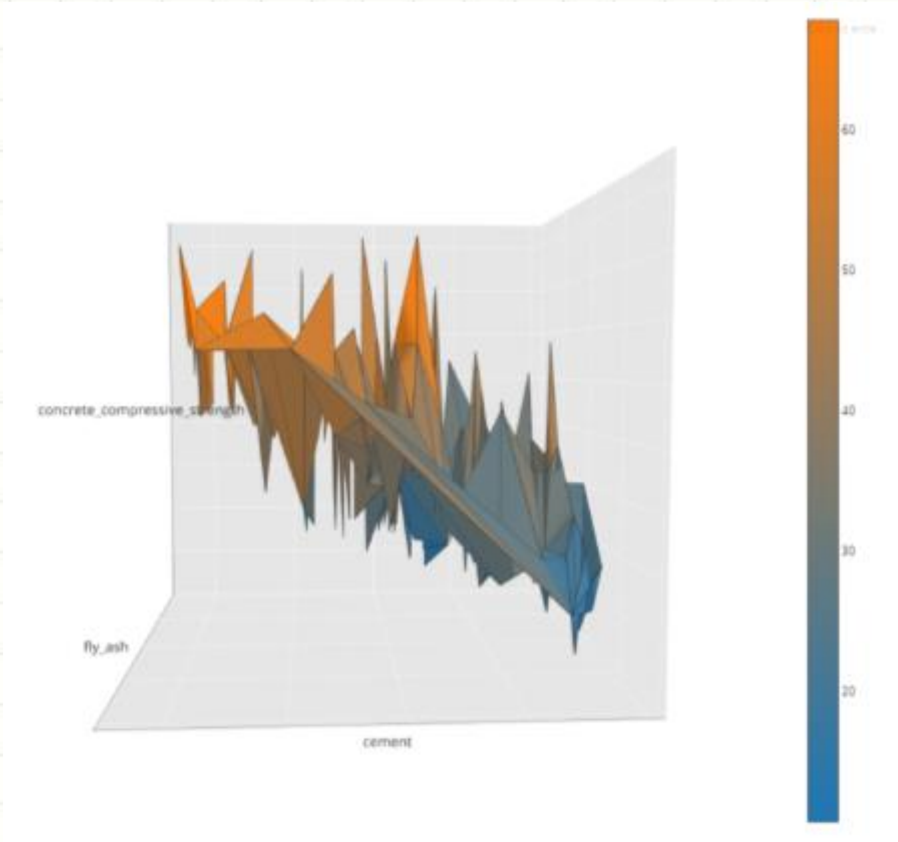
第五部分：資料內容

	cement	blast_furnace_slag	fly_ash	water	superplasticizer	coarse_aggregate	fine_aggregate	age	concrete_compressive_strength
0	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.99
1	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.89
2	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.27
3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.05
4	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.30
5	266.0	114.0	0.0	228.0	0.0	932.0	670.0	90	47.03
6	380.0	95.0	0.0	228.0	0.0	932.0	594.0	365	43.70
7	380.0	95.0	0.0	228.0	0.0	932.0	594.0	28	36.45
8	266.0	114.0	0.0	228.0	0.0	932.0	670.0	28	45.85
9	475.0	0.0	0.0	228.0	0.0	932.0	594.0	28	39.29

目標：以plotly套件繪製常用之統計圖表。

圖表

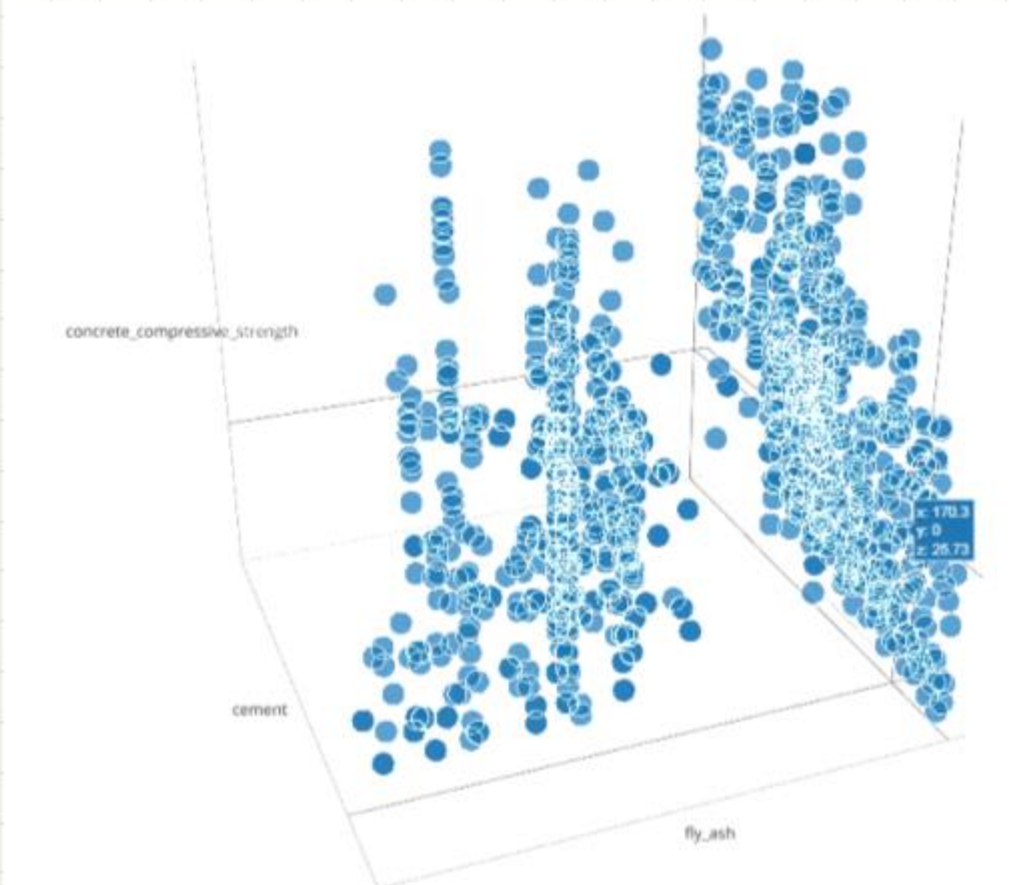
3D圖



將水泥數據中cement及fly_ash做xy軸，
對concrete_compressive_strength做圖
互動網頁版請參閱https://plot.ly/~sigmaplot/13.embed

圖表

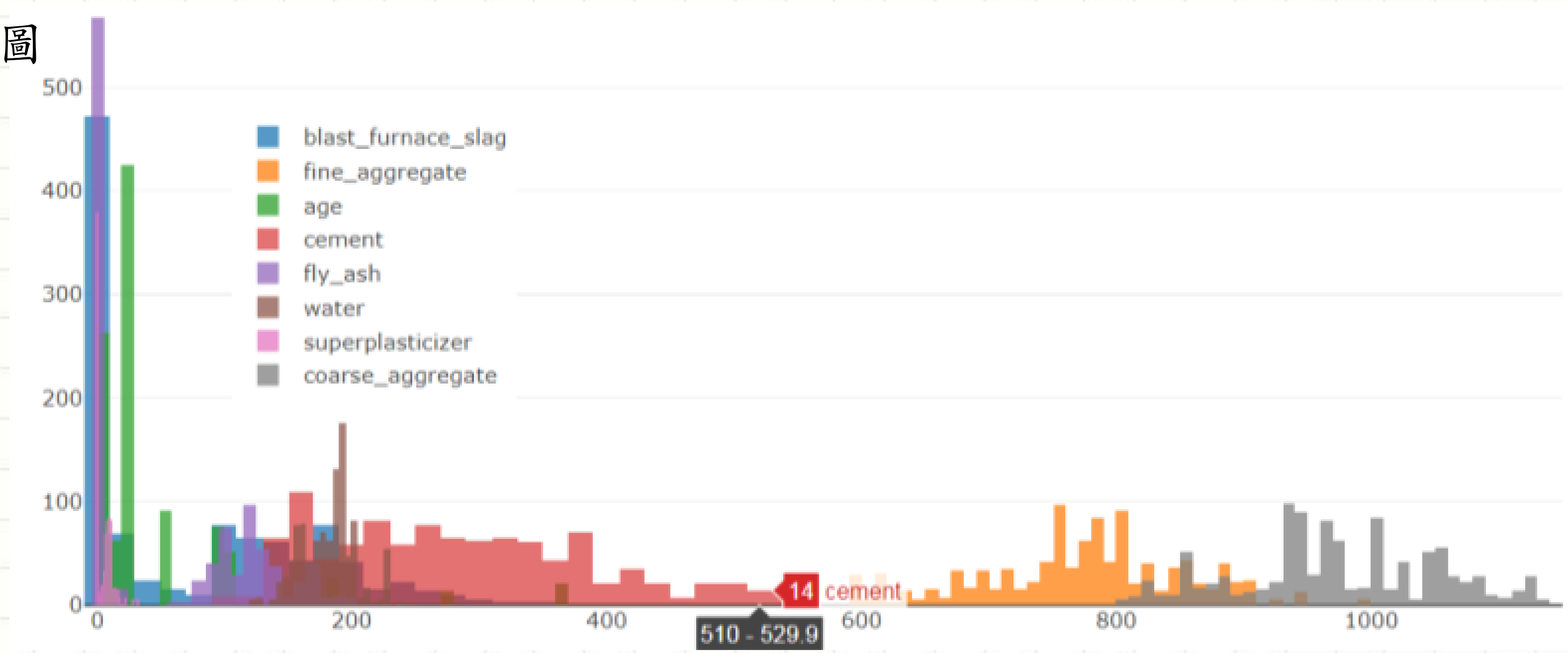
3D散佈圖



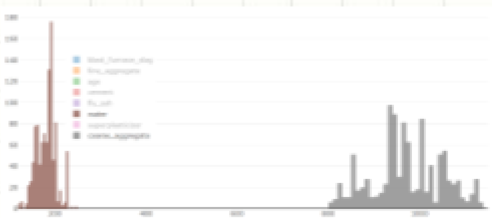
將水泥數據中cement及fly_ash做xy軸，
對concrete_compressive_strength做點散佈圖
互動網頁版請參閱https://plot.ly/~sigmaplot/23.embed

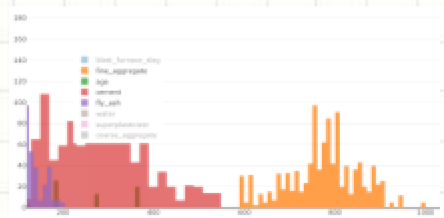
圖表

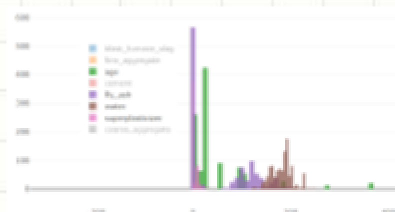
直方圖



將水泥數據中個參數畫成直方圖，可在圖例中點選要選擇之數據
如下圖只顯示部分類別的數據



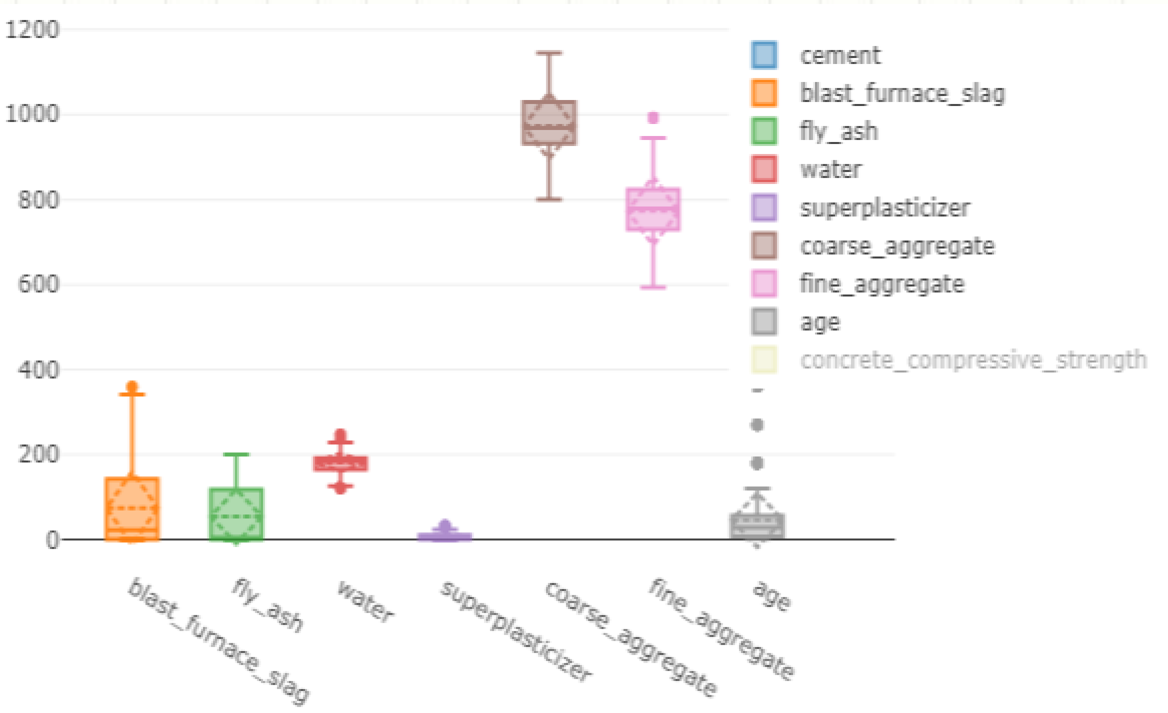


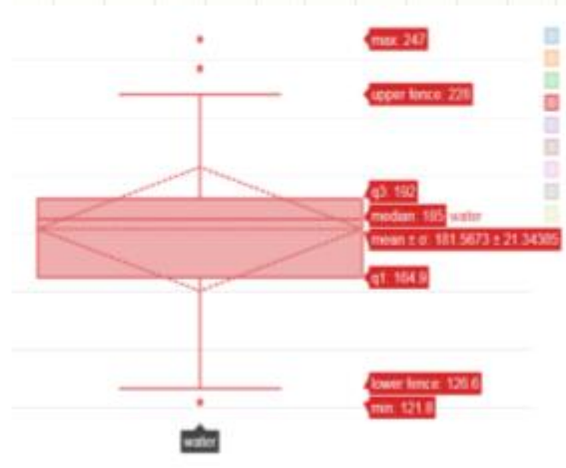


互動網頁版請參閱https://plot.ly/~sigmaplot/19.embed

圖表

boxplot



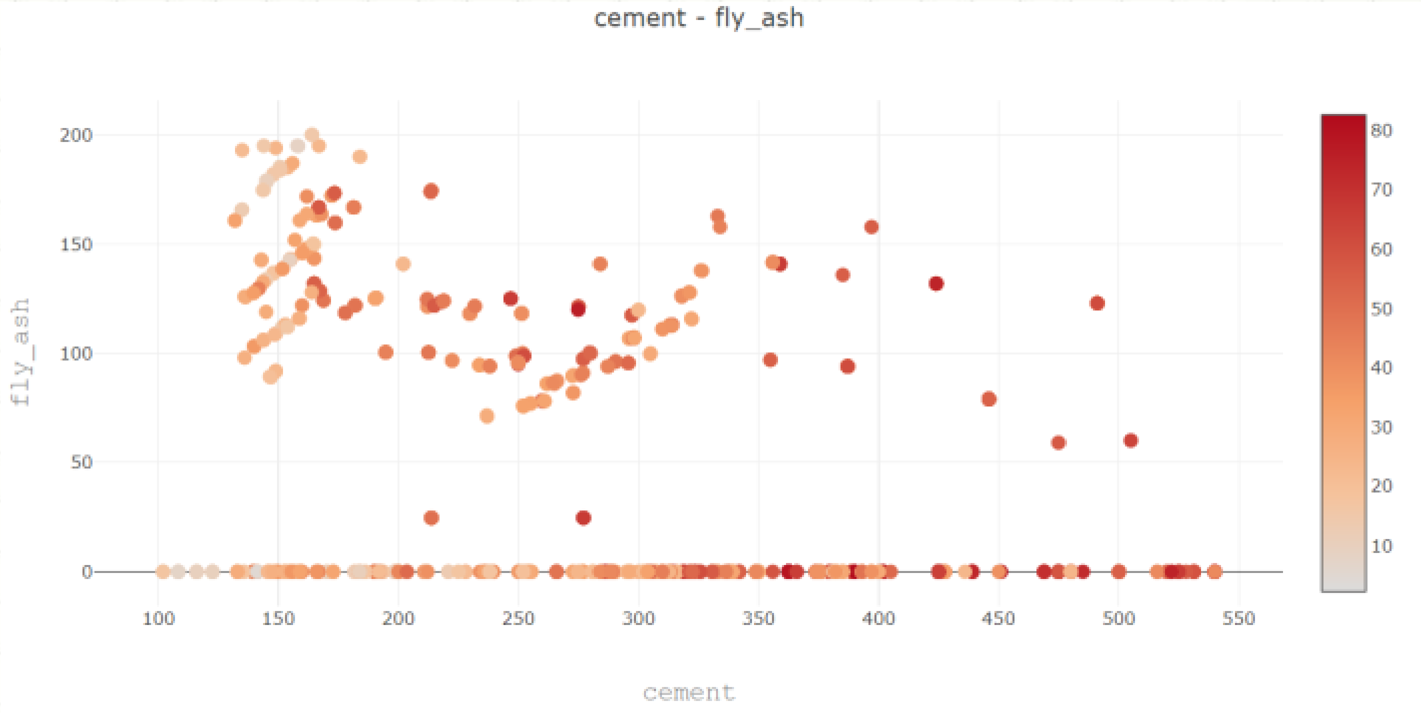


可以透過此圖了解數據集各變數基本敘述統計資料及約略資料分佈，並且可以看出離群值。

互動網頁版請參閱https://plot.ly/~sigmaplot/27.embed

圖表

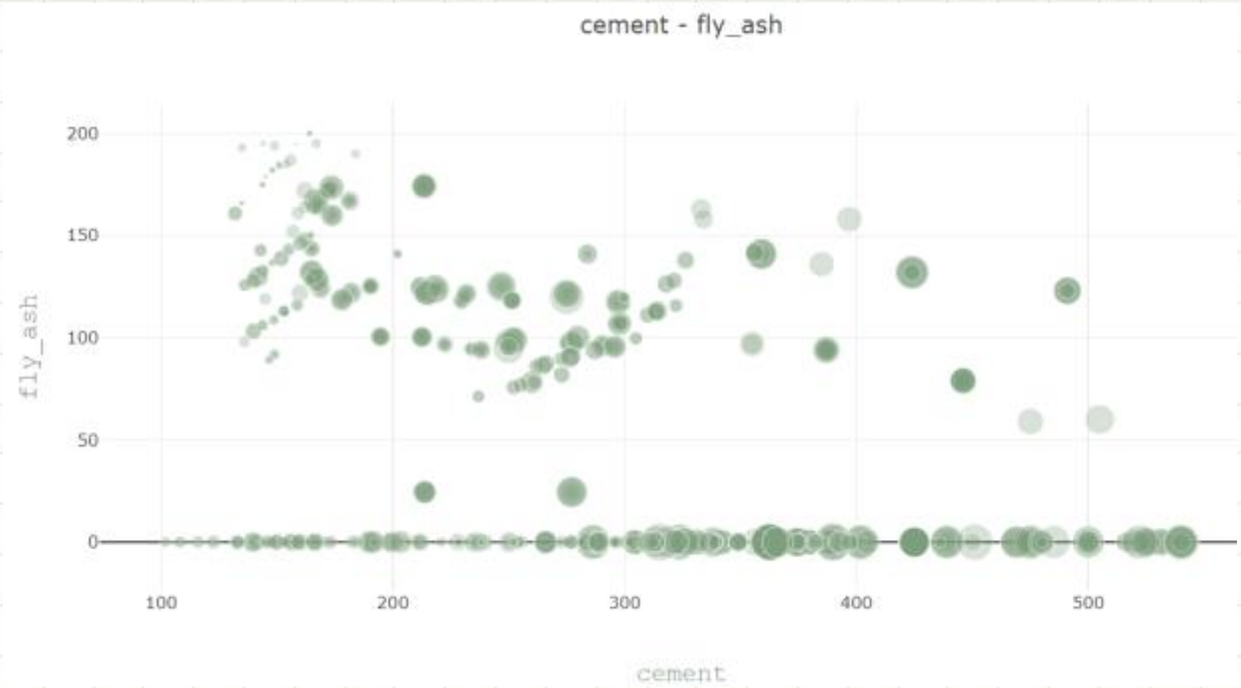
散佈圖



以cement及fly_ash做散佈圖，並以concrete_compressive_strength。
之數值大小以顏色深淺做表示。
互動網頁版請參閱https://plot.ly/~sigmaplot/31.embed

圖表

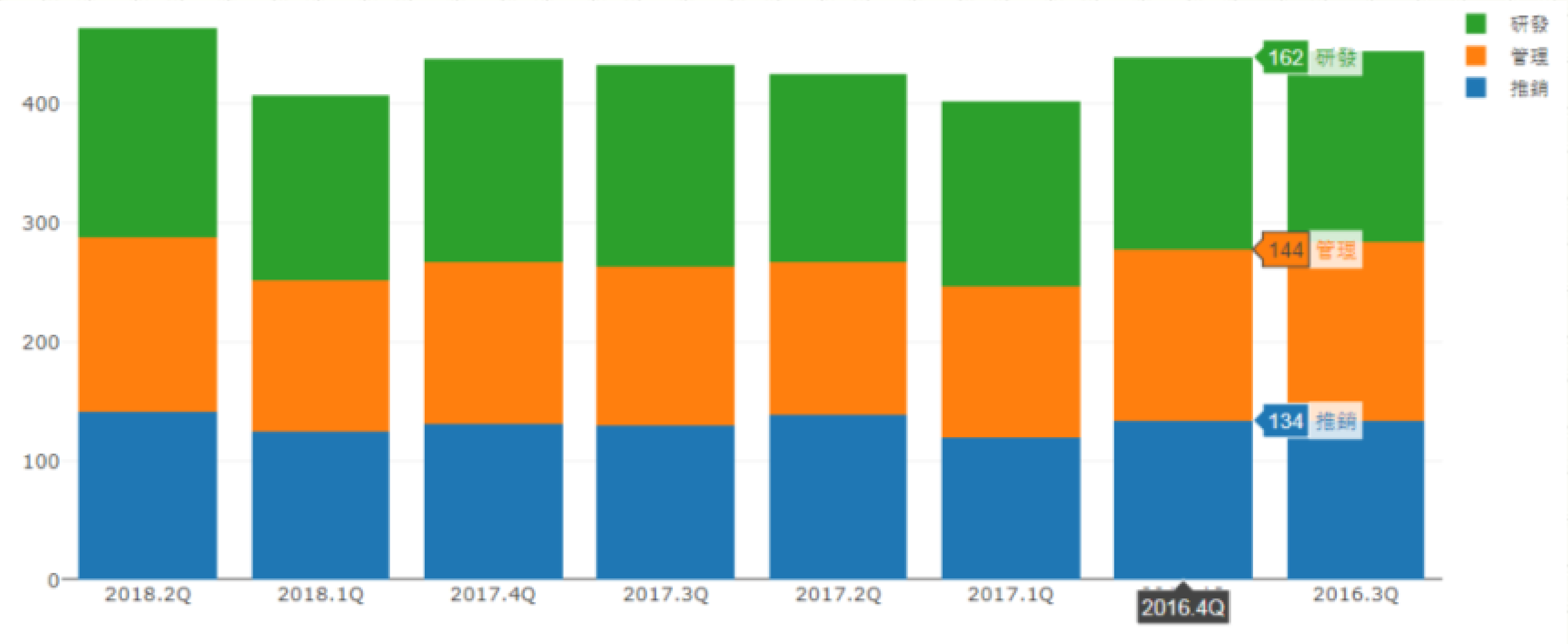
散佈圖



以cement及fly_ash做散佈圖，並以concrete_compressive_strength。之數值大小以元圈大小做表示。
互動網頁版請參閱https://plot.ly/~sigmaplot/35.embed

圖表

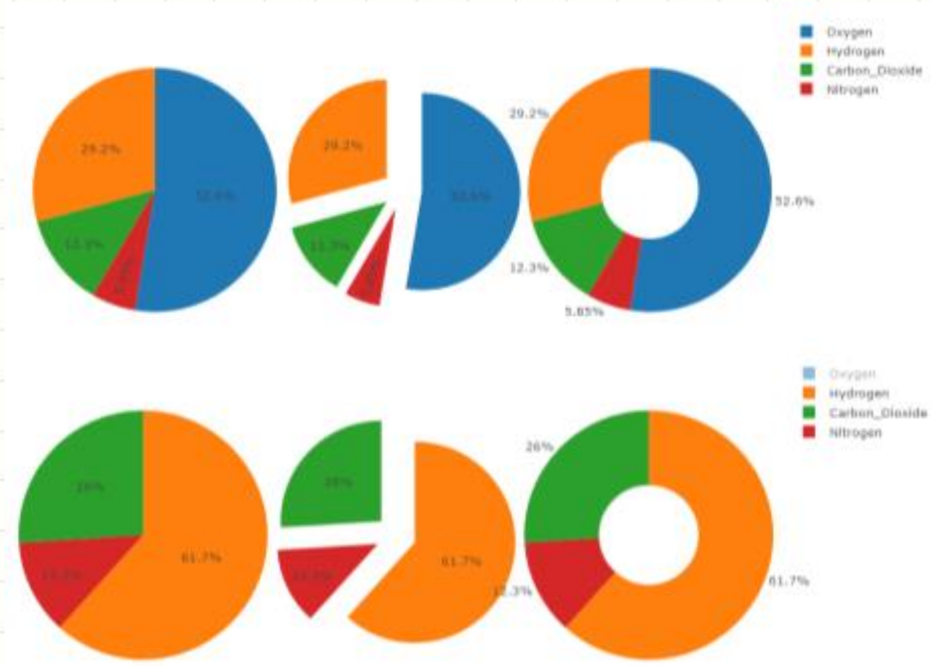
長條圖



對不同季的營業費用做成長條圖，可透過點選圖表看到詳細數字，在圖例可以選擇該組數據是否顯示。
互動網頁版請參閱https://plot.ly/~sigmaplot/41.embed

圖表

圖餅圖

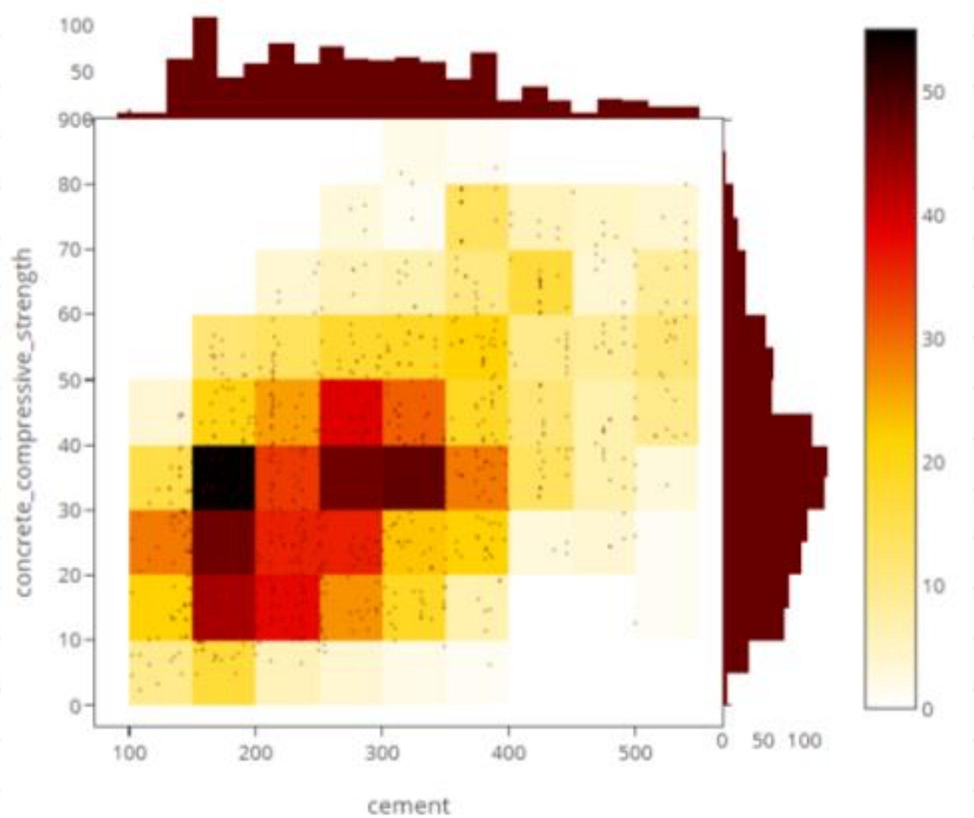


不同格式的圓餅圖展示，可透過點選圖表看到詳細數字，在圖例可以選擇該組數據是否顯示。

互動網頁版請參閱https://plot.ly/~sigmaplot/45.embed

圖表

2D密度分布圖



兩組數據除了用長條圖表示，其中的關聯性用中間分布圖表示，
顏色越深表試關聯性越大，可藉由此圖看出關聯性的差異。
互動網頁版請參閱https://plot.ly/~sigmaplot/49.embed

第六部分：資料內容

Index	authors	category	date	headline	link	short_description	text	words
0	Melissa Jeltsen	CRIME	2018-05-26 00:00:00	There Were 2 Mass Shootin…	https:// www.huffingt…	She left her husband. He …	There Were 2 Mass Shootin…	[87, 95, 260, 917, 2154, 6…
1	Andy McDonald	ENTERTAINMENT	2018-05-26 00:00:00	Will Smith Joins Diplo …	https:// www.huffingt…	Of course it has a song.	Will Smith Joins Diplo …	[34, 1516, 2197, 20046,…
2	Ron Dicker	ENTERTAINMENT	2018-05-26 00:00:00	Hugh Grant Marries For …	https:// www.huffingt…	The actor and his longtime…	Hugh Grant Marries For …	[5201, 5146, 8954, 8, 1, …
3	Ron Dicker	ENTERTAINMENT	2018-05-26 00:00:00	Jim Carrey Blasts 'Cast…	https:// www.huffingt…	The actor gives Dems a…	Jim Carrey Blasts 'Cast…	[2198, 9428, 2458, 47694,…
4	Ron Dicker	ENTERTAINMENT	2018-05-26 00:00:00	Julianna Margulies Us…	https:// www.huffingt…	The "Dietland" a…	Julianna Margulies Us…	[36179, 26511, 1605,…
5	Ron Dicker	ENTERTAINMENT	2018-05-26 00:00:00	Morgan Freeman 'Dev…	https:// www.huffingt…	"It is not right to equ…	Morgan Freeman 'Dev…	[3894, 11482, 20047, 10, 2…
6	Ron Dicker	ENTERTAINMENT	2018-05-26 00:00:00	Donald Trump Is Lovin' Ne…	https:// www.huffingt…	It's catchy, all right.	Donald Trump Is Lovin' Ne…	[55, 20, 7, 14367, 27, 3…
7	Todd Van Luling	ENTERTAINMENT	2018-05-26 00:00:00	What To Watch On Amazon Pr…	https:// www.huffingt…	There's a great mini-s…	What To Watch On Amazon Pr…	[33, 2, 178, 9, 1839, 122…
8	Andy McDonald	ENTERTAINMENT	2018-05-26 00:00:00	Mike Myers Reveals He'd…	https:// www.huffingt…	Myer's kids may be pushi…	Mike Myers Reveals He'd…	[735, 11483, 775, 2459, 9…
9	Todd Van Luling	ENTERTAINMENT	2018-05-26 00:00:00	What To Watch On Hulu That…	https:// www.huffingt…	You're getting a re…	What To Watch On Hulu That…	[33, 2, 178, 9, 8745, 156…
10	Sebastian Murdock	ENTERTAINMENT	2018-05-26 00:00:00	Justin Timberlake V…	https:// www.huffingt…	The pop star also wore a …	Justin Timberlake V…	[1331, 5377, 3831, 453, 1…
11		WORLD NEWS	2018-05-26 00:00:00	South Korean President Me…	https:// www.huffingt…	The two met to pave the …	South Korean President Me…	[430, 2128, 72, 2284, 28…

文本之類別

以文本之內容
預測其類型

方法簡述

將資料集整理成適當之X及Y的模式做訓練

→

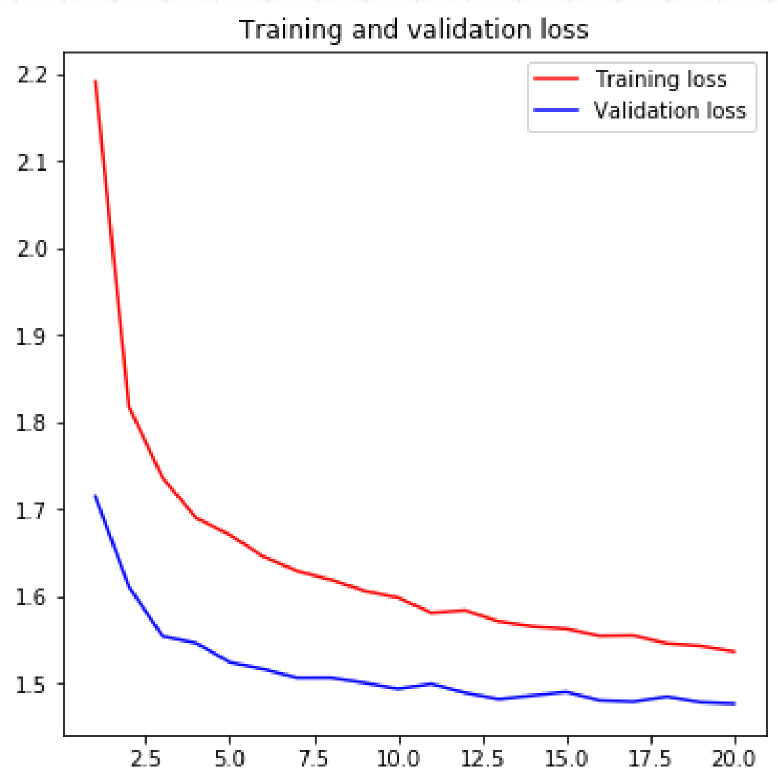
運用TextCNN、Bidirectional GRU及AttentionLSTM等模型，將數據帶入訓練及測試

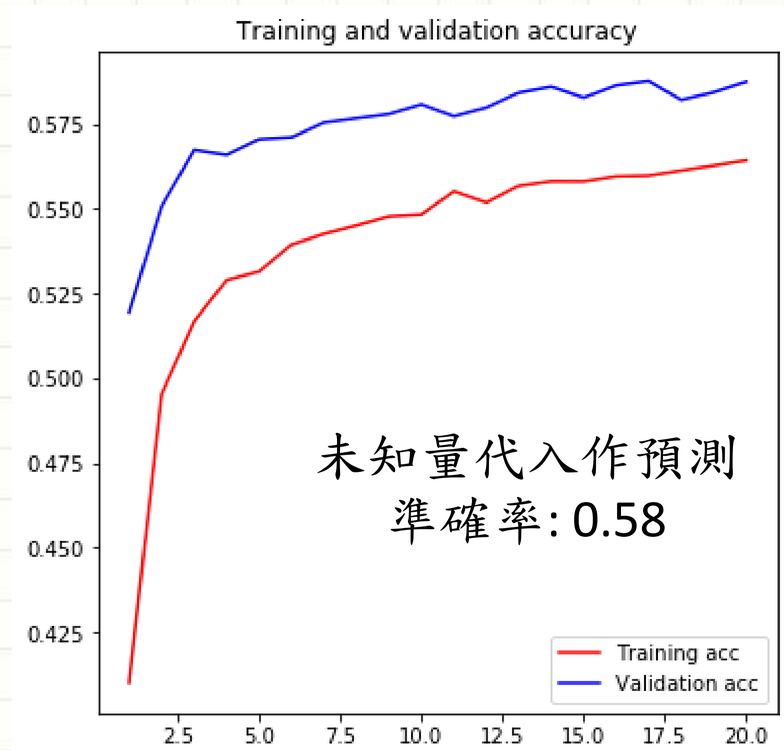
→

比較不同方法所得預測值與實際值之差異

預測結果

TextCNN

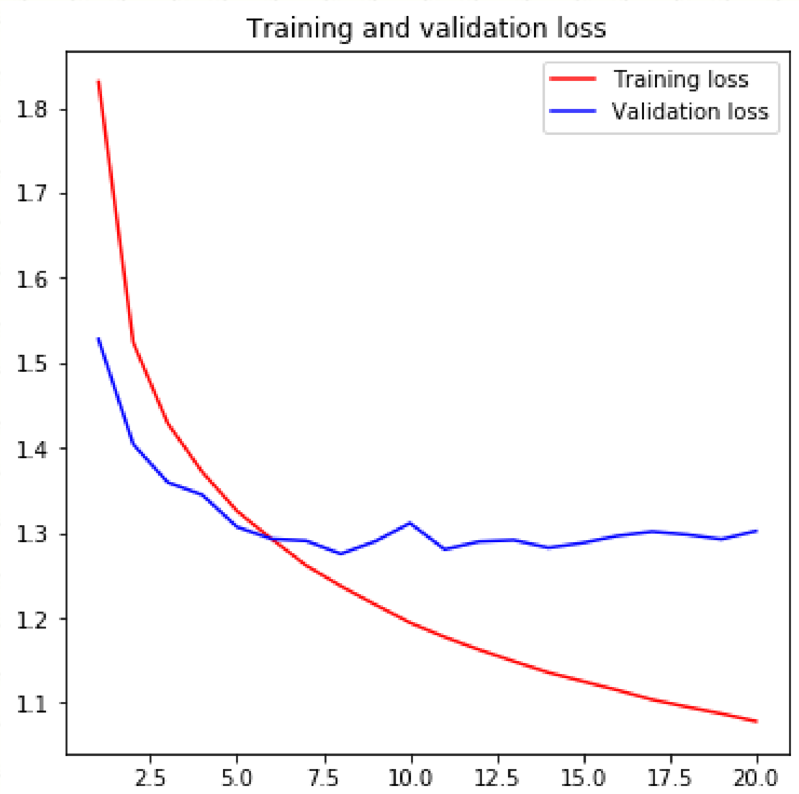


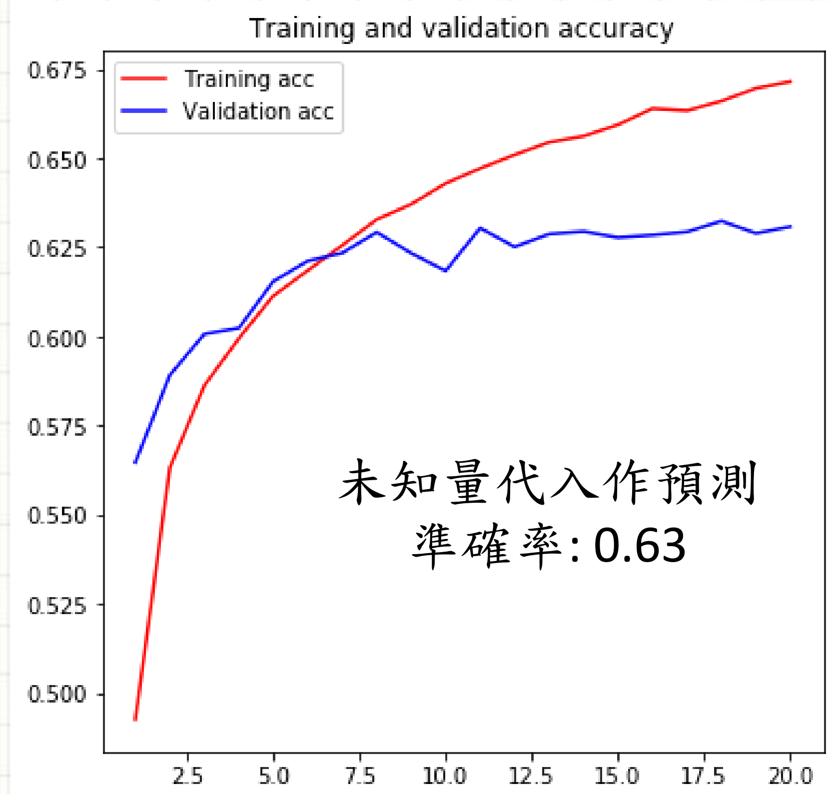


由loss圖表中，可以看出在訓練模型期間並未發生overfitting。

預測結果

Bidirectional GRU

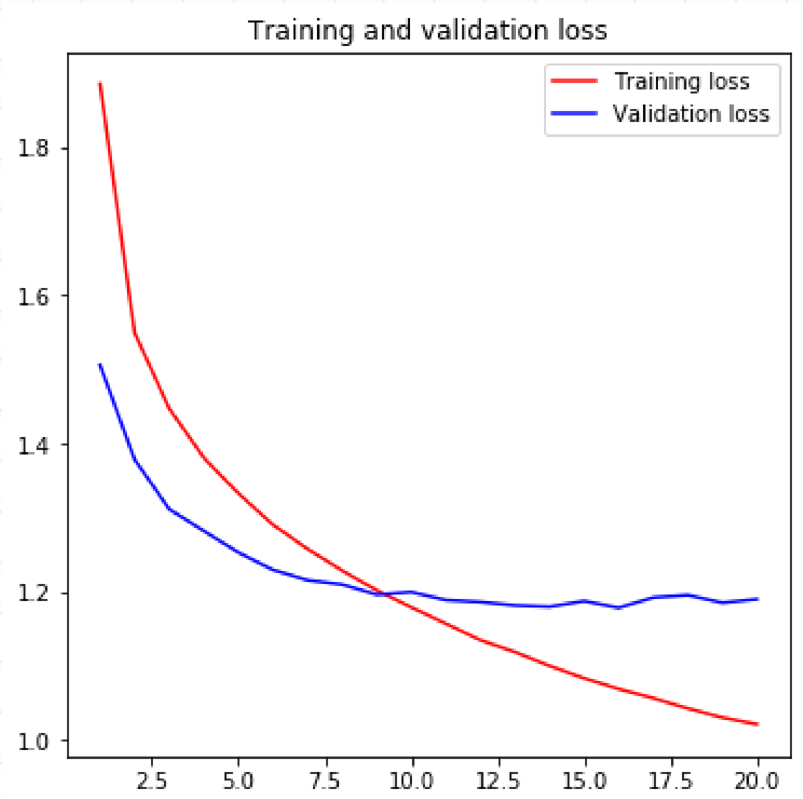


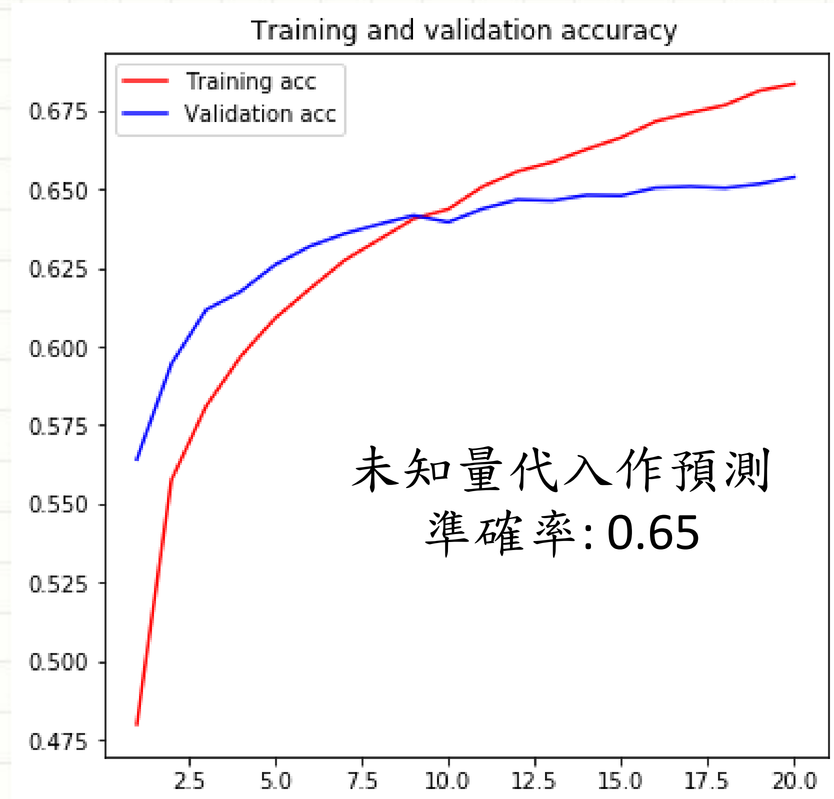


由loss圖表中，可以看出在訓練模型期間並未發生overfitting。

預測結果

AttentionLSTM





未知量代入作預測
準確率: 0.65

由loss圖表中，可以看出在訓練模型期間並未發生overfitting。

第七部分：資料內容

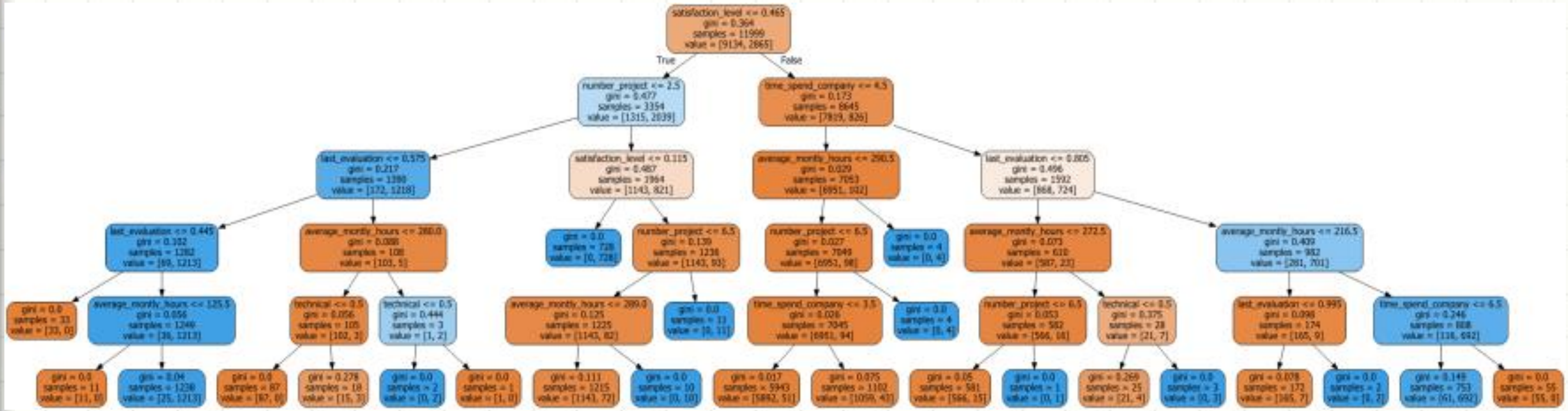
	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	sales	salary
0	0.38	0.53	2	157	3	0	1	0	sales	low
1	0.80	0.86	5	262	6	0	1	0	sales	medium
2	0.11	0.88	7	272	4	0	1	0	sales	medium
3	0.72	0.87	5	223	5	0	1	0	sales	low
4	0.37	0.52	2	159	3	0	1	0	sales	low
5	0.41	0.50	2	153	3	0	1	0	sales	low
6	0.10	0.77	6	247	4	0	1	0	sales	low
7	0.92	0.85	5	259	5	0	1	0	sales	low
8	0.89	1.00	5	224	5	0	1	0	sales	low
9	0.42	0.53	2	142	3	0	1	0	sales	low
10	0.45	0.54	2	135	3	0	1	0	sales	low
11	0.11	0.81	6	305	4	0	1	0	sales	low
12	0.84	0.92	4	234	5	0	1	0	sales	low
13	0.41	0.55	2	148	3	0	1	0	sales	low
14	0.36	0.56	2	137	3	0	1	0	sales	low
15	0.38	0.54	2	143	3	0	1	0	sales	low
16	0.45	0.47	2	160	3	0	1	0	sales	low
17	0.78	0.99	4	255	6	0	1	0	sales	low
18	0.45	0.51	2	160	3	1	1	1	sales	low
19	0.76	0.89	5	262	5	0	1	0	sales	low

此參數為離職與未離職，是要預測之項目。

以這些參數建構預測模型

第七部分：資料內容

以決策樹建構預測模型:



	實際在職	實際離職
預測在職	2,263	31
預測離職	55	651

準確率:**0.97**，可以由決策樹圖了解各參數影響離職的重要性，例如:滿意度的高低在此例為最重要，次重要的是當滿意度<0.465，可依照做專案數量做分類，而若當滿意度>0.465，可依照待公司之時間長短做分類，可依此方式了解其中因果關係。