學號:0411509　姓名:許家維

採用數據:　電網穩定性模擬數據

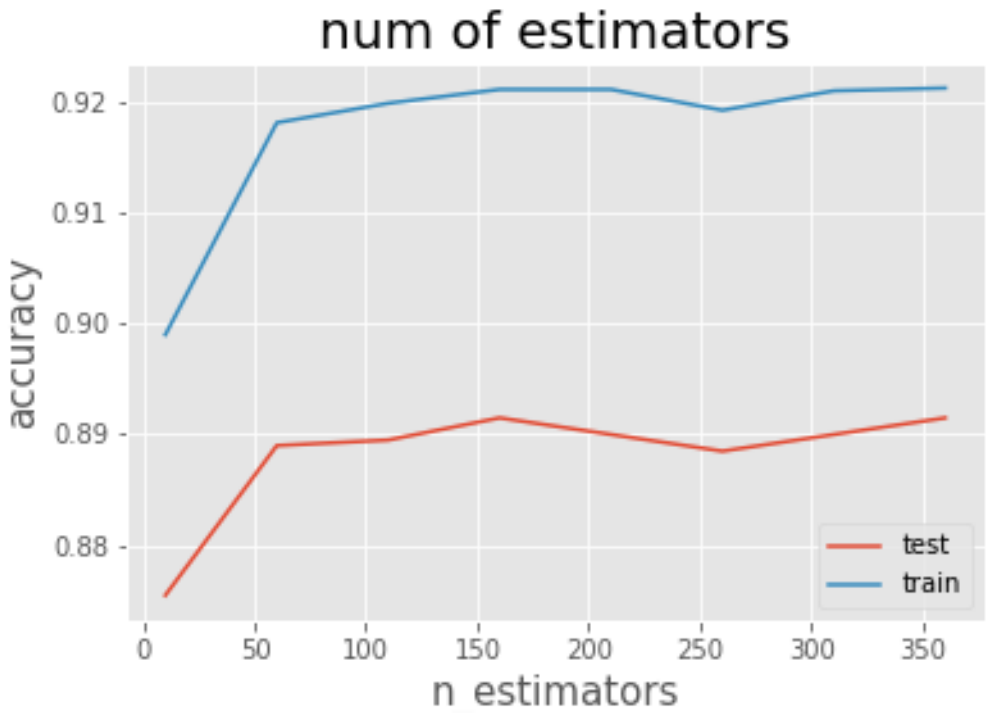
其中包含了(消耗功率、生產電力、反應時間等等)，並預測分類該電網是否穩定。

數據來源:

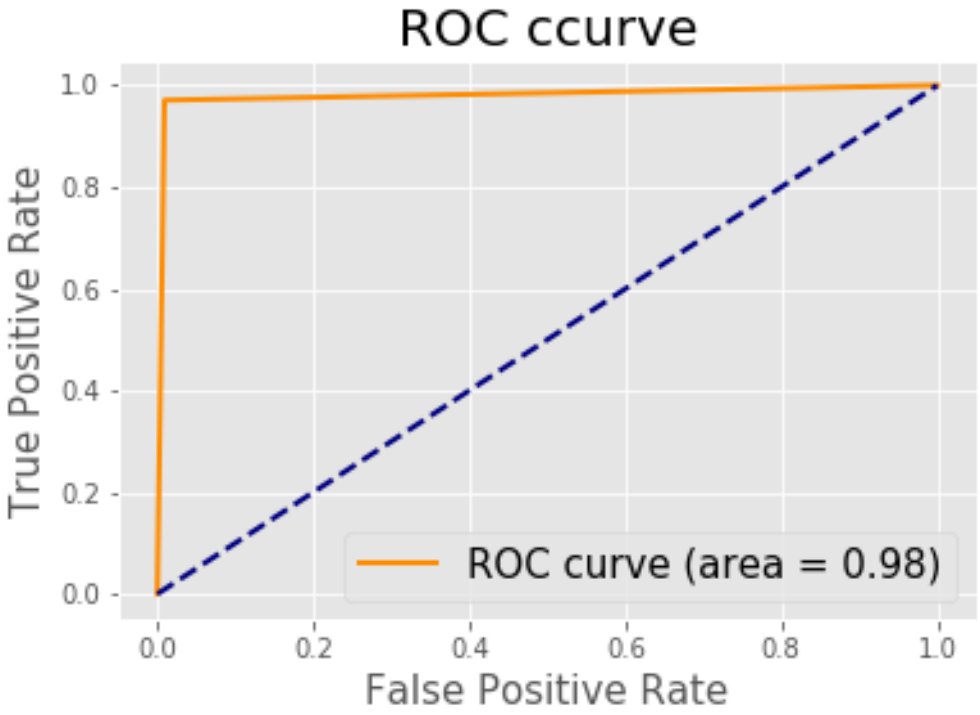
http://archive.ics.uci.edu/ml/datasets/Electrical+Grid+Stability+Simulted+Data+#

分析過程

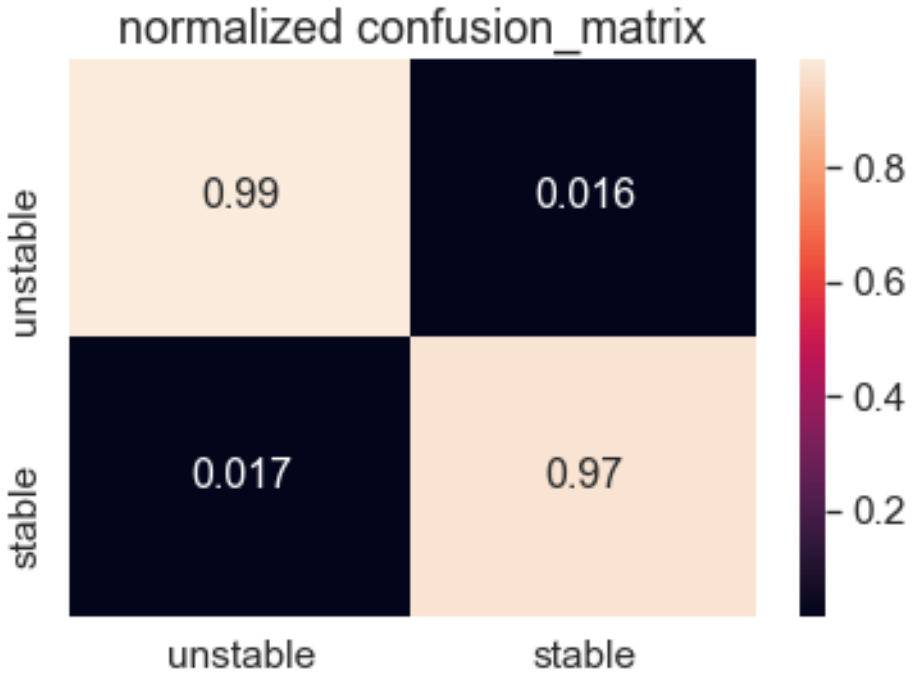
此次數據採用 Random Forest 分類，其中是利用 ensembl 多個決策樹達成分類預測的，其中若其中樹木數量是一個很重要的超參數，太多容易造成過擬和(overfitting)，過低則會有欠擬和(under fitting)，透過逐一分析，可發現在 n=150 下有較良好的泛化能力並不失去預測準確度，在不失去 testing data 準度下提高 training data 準確度。



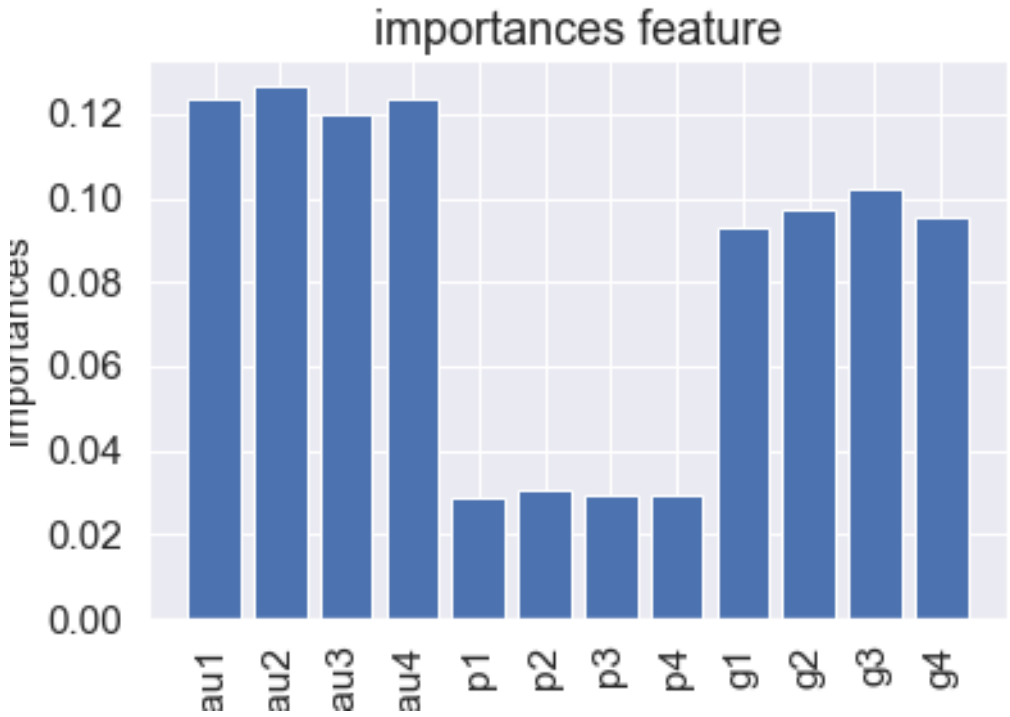
為了評量模型預測能力，我使用 Receiver operating characteristic(ROC)並計算 AUC，發現該隨機森林分類能力不錯，曲線下面積高達 0.98。



並把混淆矩陣(confusion matrix)可視化，發現雖然分類穩定的效果稍弱，但準確性也高達 97%

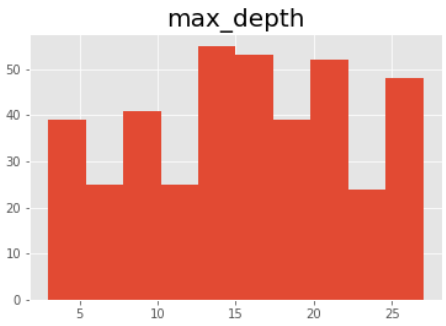


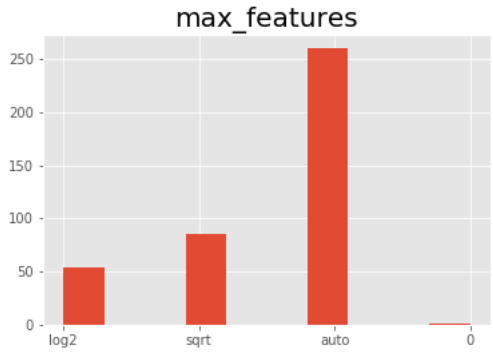
其中特過此隨機森林的模型也可以得知何種參數對於分類的影響性，透過分析葉子與根的熵(entropy)，即可進一步對數據的解讀，可以觀察到 tau(reaction time of participant)和 g(coefficient　proportional to price elasticity)對於分類電網是否穩定很重要。

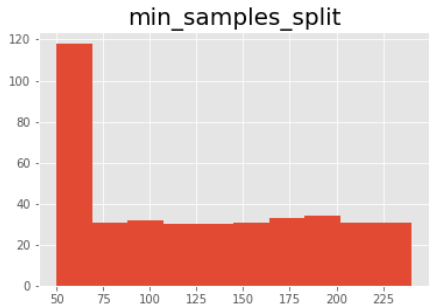


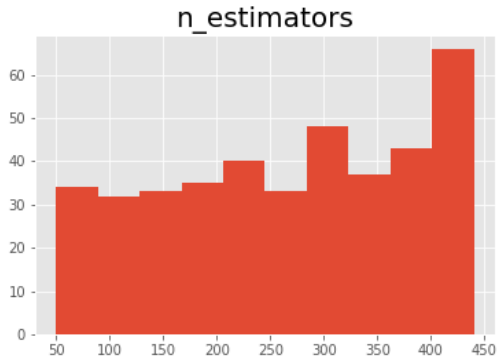
為了更加挖掘隨機森林潛力，我透過 Bayesian Hyperparam Optimization 自動調整超參數，包含了 criterio、n_estimators、max_depth 等，其透過高斯過程自動找尋最佳超參數，畢竟由於超參數眾多無法使用棋盤法一一尋找。

以下是我疊代 400 次後呈現的超參數分佈:









整理出最好的超參數:

criterion: gini

max_depth: 13

max_features: auto

min_samples_split:50

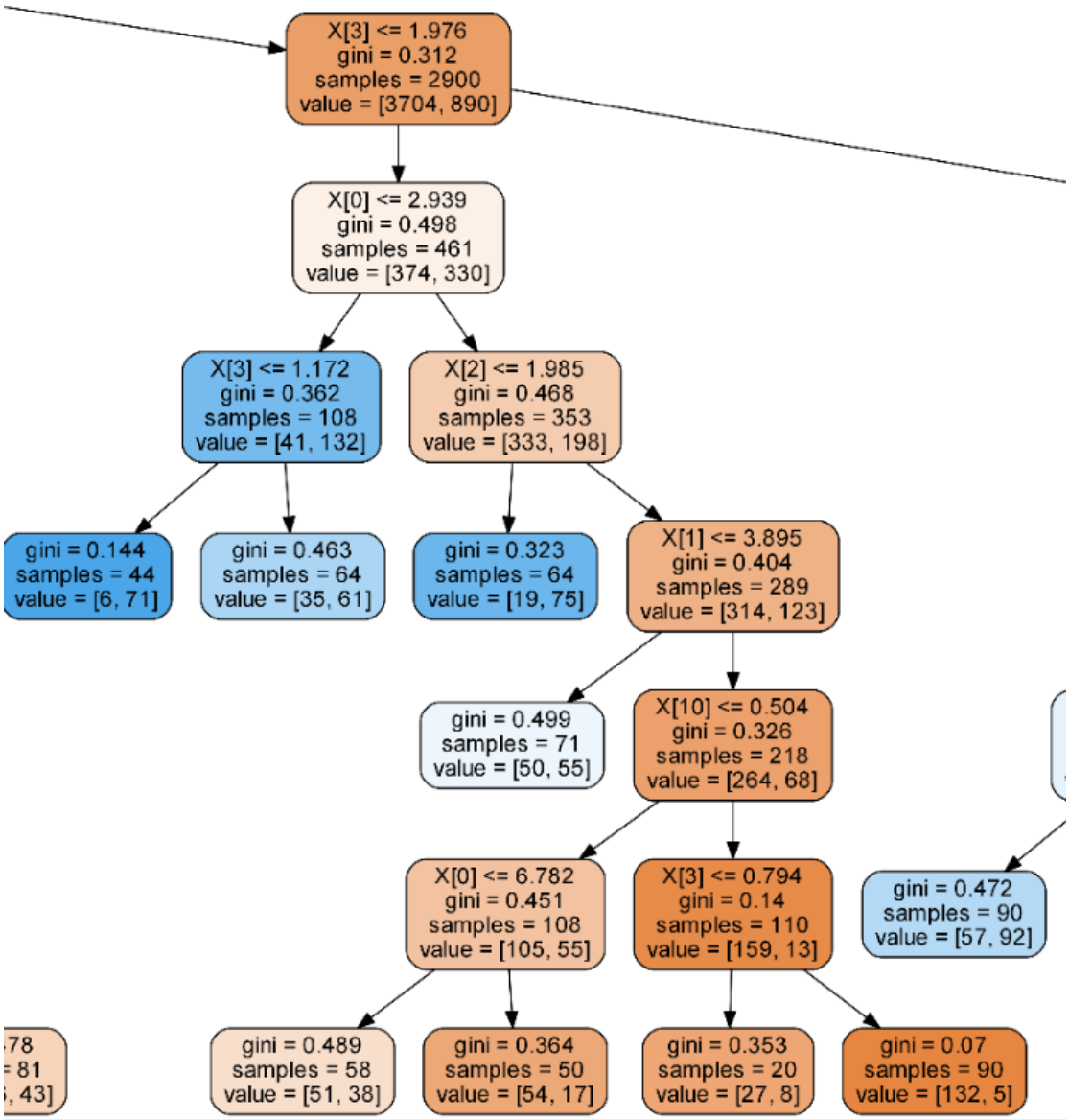
n_estimators:430

總結

透過隨機森林的算法進行分類，並計算出分類結果的精準度、召回率(recall)、f1-score 可以得到以下總表，可以顯示該分類方法十分優異，對於class0(unstable)和 class1(stable)的分類都有著極高的正確率。

	precision	recall	f1-score	support
0.0	0.98	0.99	0.99	6380
1.0	0.98	0.97	0.98	3620
micro avg	0.98	0.98	0.98	10000
macro avg	0.98	0.98	0.98	10000
weighted avg	0.98	0.98	0.98	10000

以下是該隨機森林下的其中一個分支的可視化


X[3] <= 1.976
gini = 0.312
samples = 2900
value = [3704, 890]
X[0] <= 2.939
gini = 0.498
samples = 461
value = [374, 330]
X[3] <= 1.172
gini = 0.362
samples = 108
value = [41, 132]
X[2] <= 1.985
gini = 0.468
samples = 353
value = [333, 198]
gini = 0.144
samples = 44
value = [6, 71]
gini = 0.463
samples = 64
value = [35, 61]
gini = 0.323
samples = 64
value = [19, 75]
X[1] <= 3.895
gini = 0.404
samples = 289
value = [314, 123]
gini = 0.499
samples = 71
value = [50, 55]
X[10] <= 0.504
gini = 0.326
samples = 218
value = [264, 68]
X[0] <= 6.782
gini = 0.451
samples = 108
value = [105, 55]
X[3] <= 0.794
gini = 0.14
samples = 110
value = [159, 13]
gini = 0.472
samples = 90
value = [57, 92]
78
= 81
, 43]
gini = 0.489
samples = 58
value = [51, 38]
gini = 0.364
samples = 50
value = [54, 17]
gini = 0.353
samples = 20
value = [27, 8]
gini = 0.07
samples = 90
value = [132, 5]