

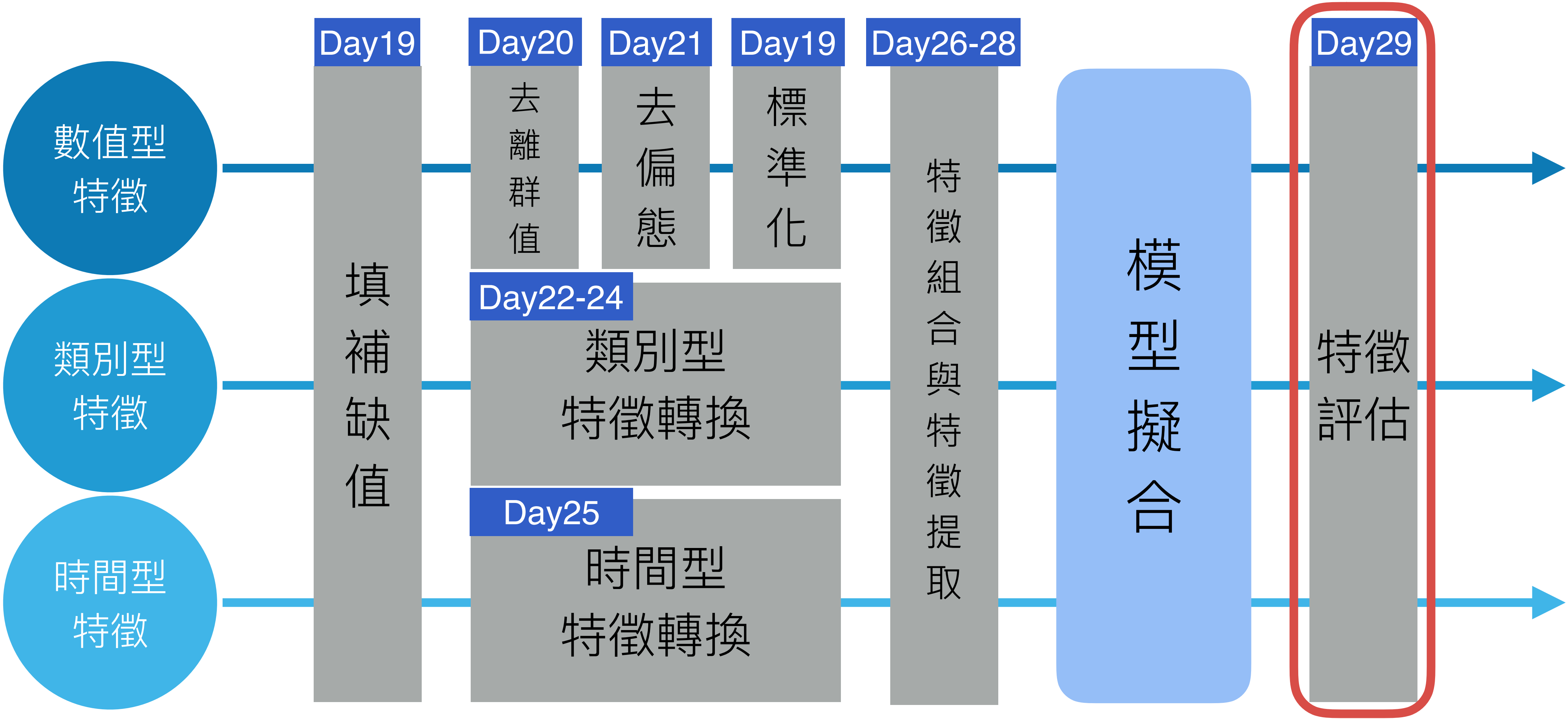
Day 29

特徵工程

特徵評估



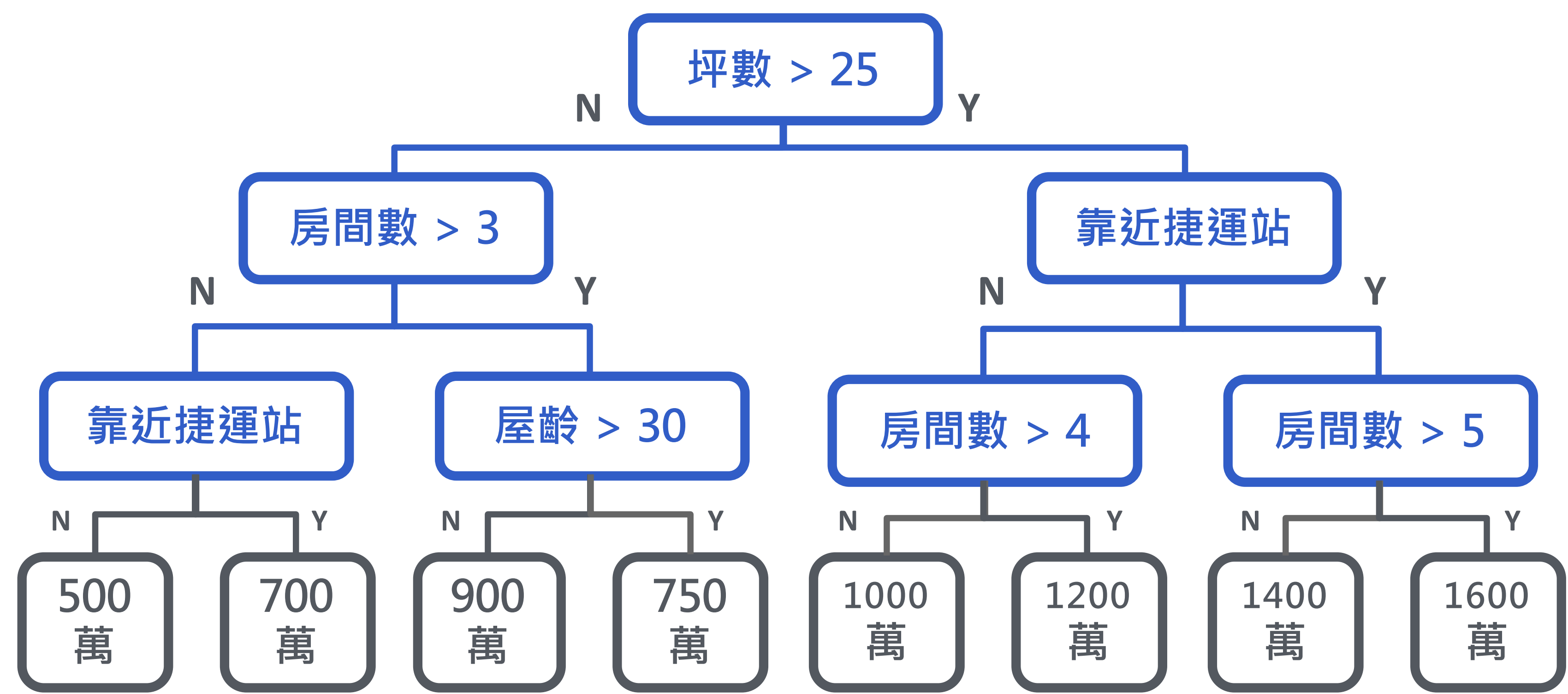
特徵工程 - 學習地圖



細說特徵重要性 (1 / 3)

讓我們先來看看什麼是特徵重要性：

下列是房價預估決策樹的預測圖，四個特徵 (坪數、房間數、屋齡、是否靠近捷運站) 之中，請問你覺得哪一個特徵比較重要？

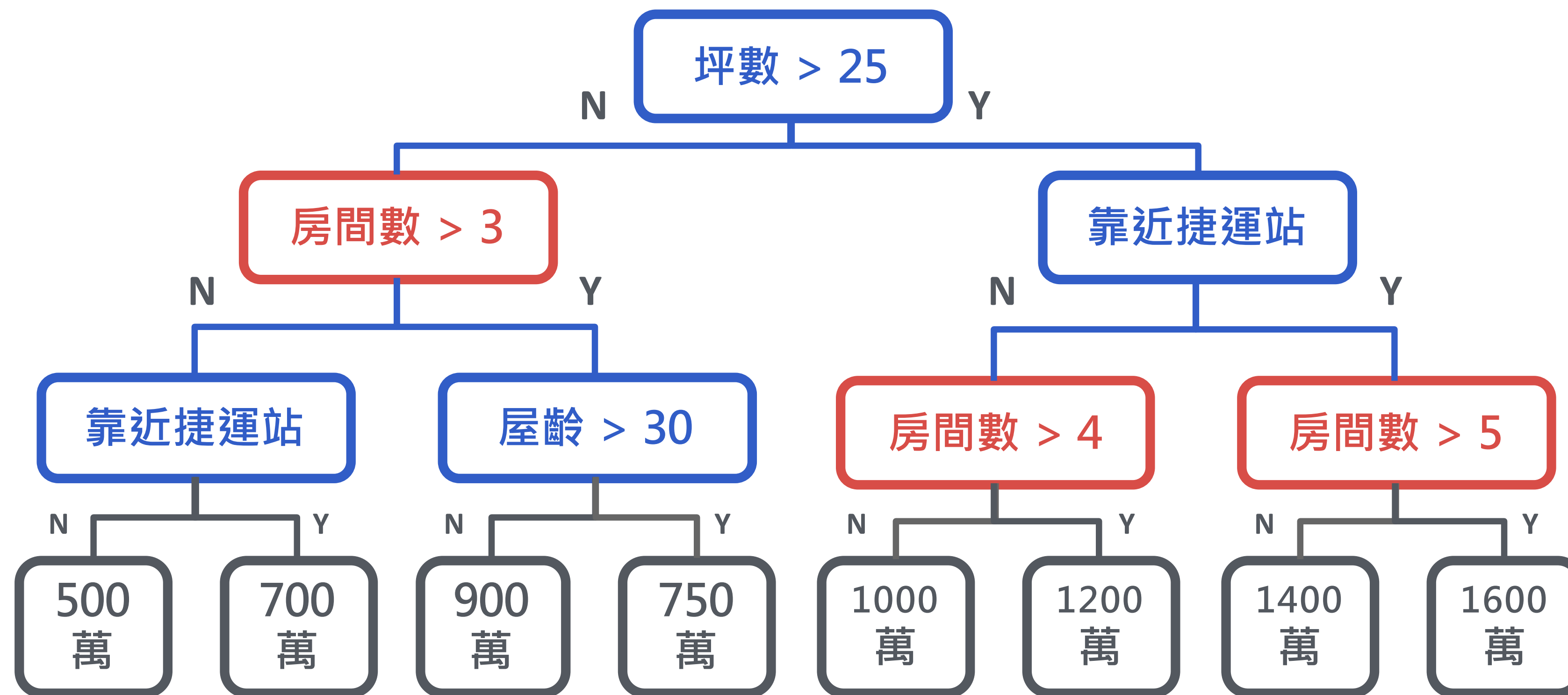


細說特徵重要性 (2 / 3)

特徵重要性預設方式是取 **特徵決定分支的次數**

此例而言：坪數x1次 房間數x3次 靠近捷運站x2次 屋齡x1次

所以最重要的特徵是 **房間數**

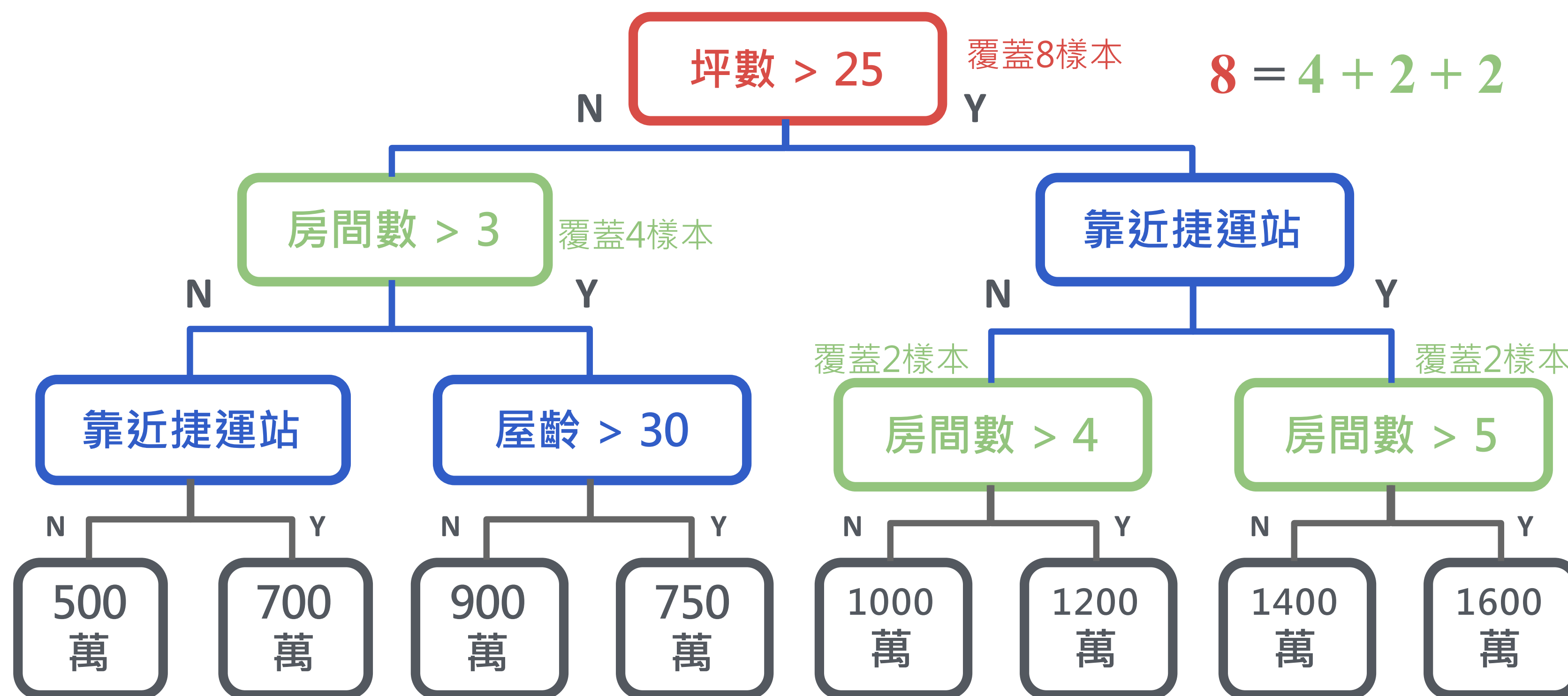


細說特徵重要性 (3 / 3)

但分支次數以外，還有兩種更直覺的特徵重要性：特徵覆蓋度、損失函數降低量

本例的特徵覆蓋度(假定八個結果樣本數量一樣多)：坪數與房間數的覆蓋度相同(都是8)

而損失函數降低量，則是要看損失函數 (loss function) 決定



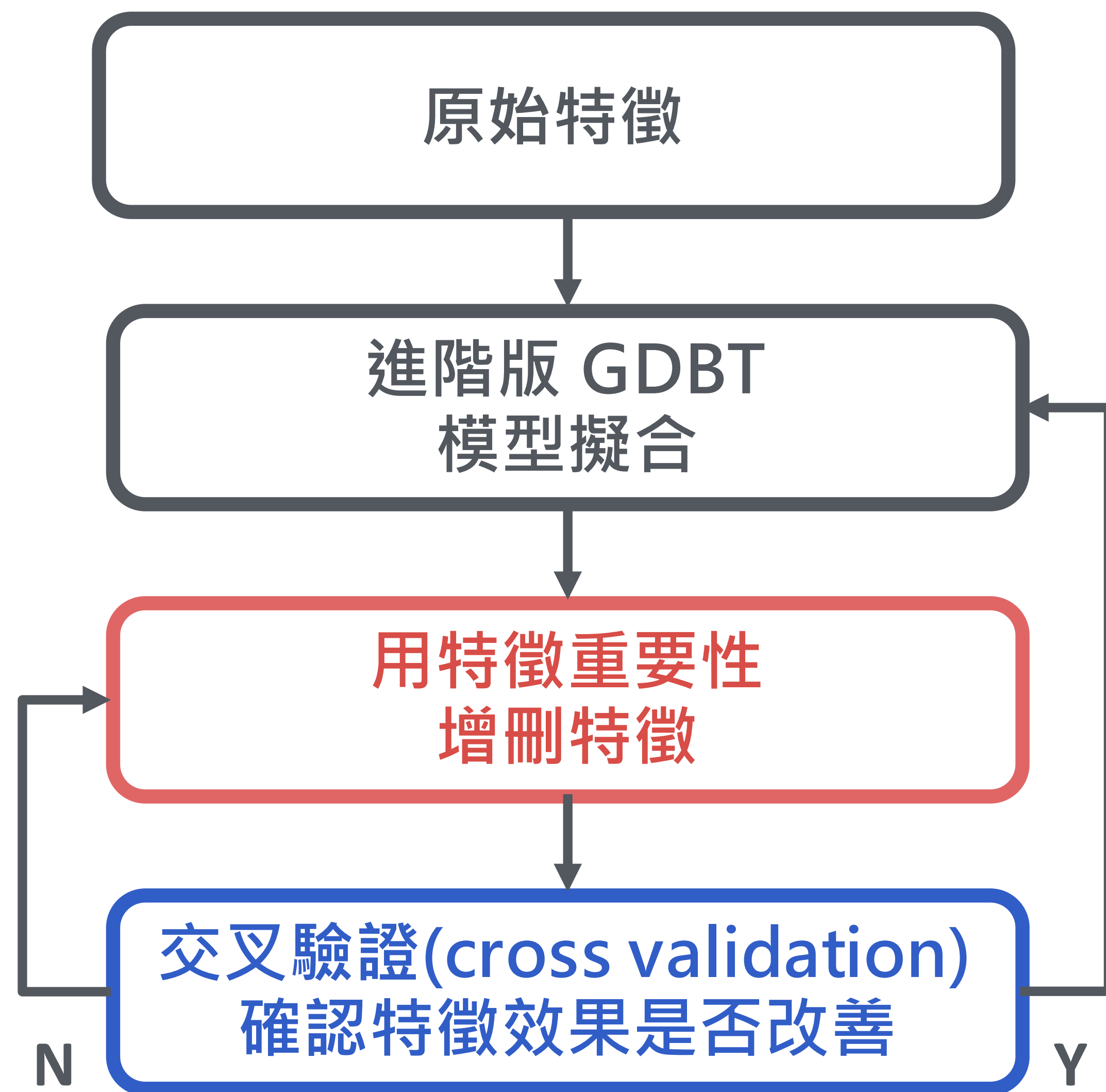
套件中的特徵重要性

- sklearn 當中的樹狀模型，都有特徵重要性這項方法
(.feature_importances_)，而實際上都是**分支次數**
- 進階版的 GDBT模型(xgboost, lightgbm, catboost) 中，
才有上述三種不同的重要性

	Xgboost 對應參數 (importance_type)	計算時間	估計精確性	sklearn 有此功能
分支次數	weight	最快	最低	O
分支覆蓋度	cover	快	中	X
損失降低量 (資訊增益度)	gain	較慢	最高	X

機器學習中的優化循環

- 機器學習特徵優化，循環方式如圖
- 其中增刪特徵指的是
 - **特徵選擇(刪除)**
 - 挑選門檻，刪除一部分特徵重要性較低的特徵
 - **特徵組合(增加)**
 - 依領域知識，對前幾名特徵做特徵組合或群聚編碼，形成更強力特徵
- 由交叉驗證確認特徵是否有改善，若沒有改善則回到上一輪重選特徵增刪
- 這樣的流程圖綜合了 **PART 2：特徵工程** 的主要內容，是這個部分的**核心知識**



排列重要性 (permutation Importance)

- 雖然特徵重要性相當實用，然而計算原理必須基於樹狀模型，於是有了可延伸至非樹狀模型的**排序重要性**
- 排序重要性計算，是打散單一特徵的資料排序順序，再用原本模型重新預測，觀察打散前後誤差會變化多少

	特徵重要性 Feature Importance	排序重要性 Permutation Importance
適用模型	限定樹狀模型	機器學習模型均可
計算原理	樹狀模型的分歧特徵	打散原始資料中單一特徵的排序
額外計算時間	較短	較長

解題時間 It's Your Turn

請跳出PDF至官網Sample Code & 作業
開始解題

