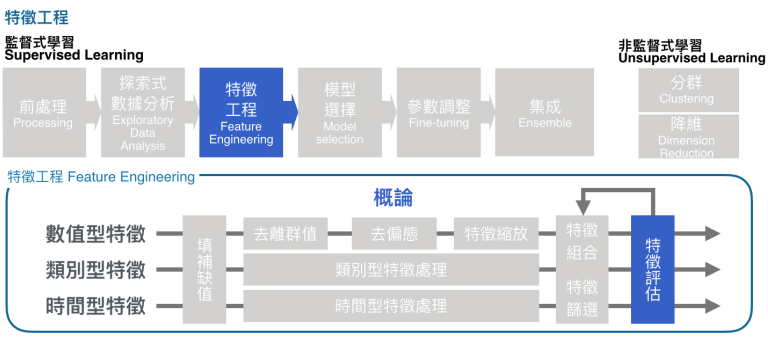
D31：特徵評估

簡報閱讀 範例與作業 問題討論

特徵評估

知識地圖



本日知識點目標

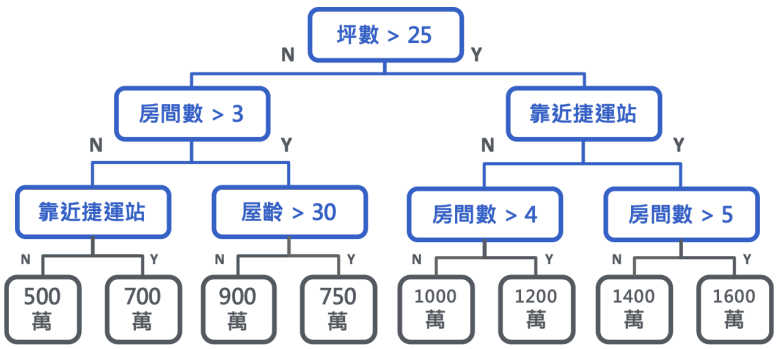
本日知識點目標

- 樹狀模型的特徵重要性，可以分為哪三種
- sklearn 樹狀模型的特徵重要性與 Xgboost 的有何不同
- 特徵工程中，特徵重要性本身的重要性是什麼

細說特徵重要性

讓我們先來看看什麼是特徵重要性：

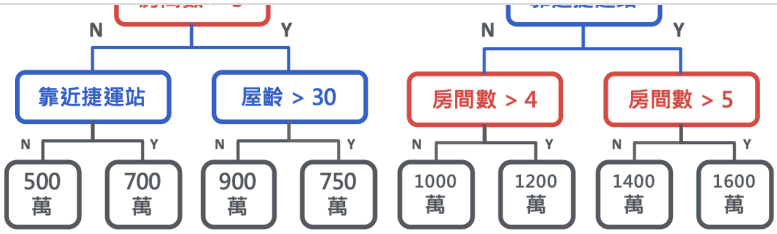
下列是房價預估決策樹的預測圖，四個特徵 (坪數、房間數、屋齡、是否靠近捷運站) 之中，請問你覺得哪一個特徵比較重要?



特徵重要性預設方式是取 特徵決定分支的次數

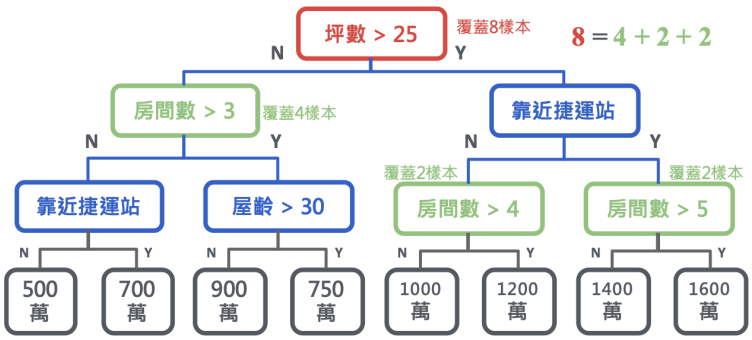
此例而言：坪數x1次 房間數x3次 靠近捷運站x2次 屋齡x1次

所以最重要的特徵是 房間數



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

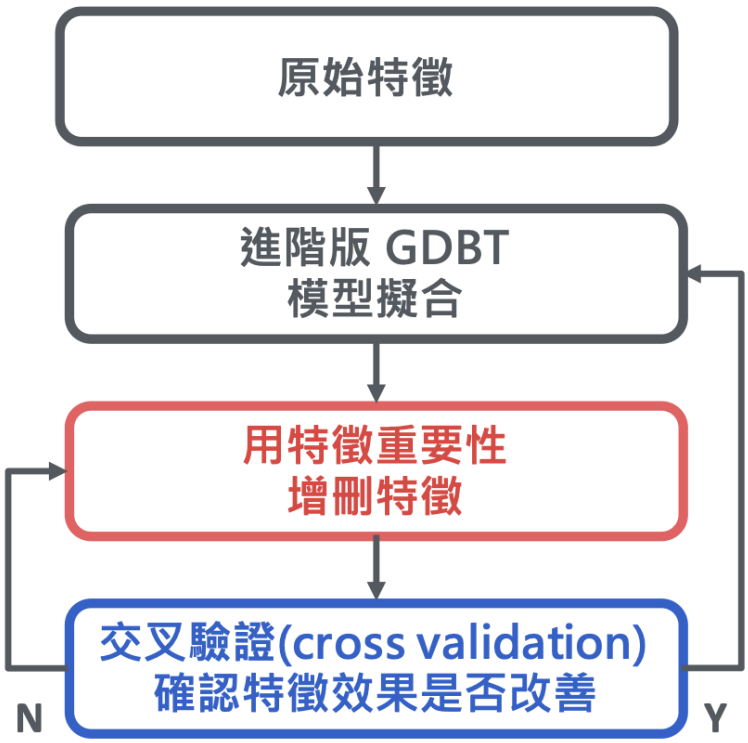
- 機器學習特徵優化，循環方式如下圖
- 其中增刪特徵指的是

  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 3 : 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

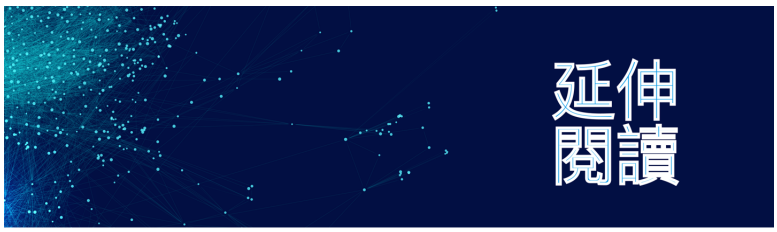
序順序，再用原本模型重新預測，觀察打散前後誤差會變化多少

	特徵重要性 Feature Impotance	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

重要知識點複習

- 樹狀模型的特徵重要性，可以分為**分支次數**、**特徵覆蓋度**、**損失函數降低量**三種
- sklearn 樹狀模型與 Xgboost 的特徵重要性，最大差異就是在 **sklearn 只有精準度最低的「分支次數」**
- 特徵重要性本身的重要性，是在於本身是**增刪特徵的重要判定準則**，在領域知識不足時，成為改善模型的最大幫手

延伸閱讀



- 除了每日知識點的基礎之外，推薦的延伸閱讀能補足學員們對該知識點的了解程度

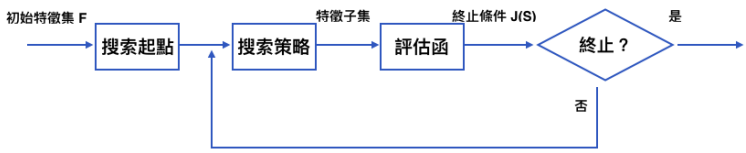
推薦延伸閱讀

機器學習 - 特徵選擇算法流程、分類、優化與發展綜述

掘金

> 機器學習 - 特徵選擇算法流
> 隨著大數據時代的到來，各行各業湧現的海量數據對數據處理的技術提出
> juejin.im

有關特徵選擇的優化流程，在這邊有更完整的說明，不過這篇文章與其說是說明，不如說更像一份索引，我們可以在這篇文章中找到相當多的名稱與論文選錄，建議同學在專題 / 競賽當中遇到瓶頸時，不妨來逛逛這篇，尋找一下靈感



Permutation Importance
Kaggle Dan B. (英文)

> Permutation Importance | Kaggle
> Explore and run machine learning code with Kaggle Notebooks | Using data from multiple data sources

- 這裡是 Kaggle 上 Dan B. 提供的課程網頁，介紹我們課程中提到的排列重要性，雖然在樹狀模型上，其精準度略遜於特徵重要性，但是這個方法在非樹狀模型上也適用，泛用性不差

容，例如 SHAP Value，能將樹狀模型預測的各個特徵影響性都可解釋化，在某些應用上，這個會比精準度還要有用

Weight	Feature
0.0750 ± 0.1159	Goal Scored
0.0625 ± 0.0791	Corners
0.0437 ± 0.0500	Distance Covered (Kms)
0.0375 ± 0.0729	On-Target
0.0375 ± 0.0468	Free Kicks
0.0187 ± 0.0306	Blocked
0.0125 ± 0.0750	Pass Accuracy %
0.0125 ± 0.0500	Yellow Card
0.0063 ± 0.0468	Saves
0.0063 ± 0.0250	Offsides
0.0063 ± 0.1741	Off-Target
0.0000 ± 0.1046	Passes
0 ± 0.0000	Red
0 ± 0.0000	Yellow & Red
0 ± 0.0000	Goals in PSO
-0.0312 ± 0.0884	Fouls Committed
-0.0375 ± 0.0919	Attempts
-0.0500 ± 0.0500	Ball Possession %

解題時間

Sample Code & 作業
開始解題

下一步：閱讀範例與完成作業