Day 28 特徵工程

特徵選擇

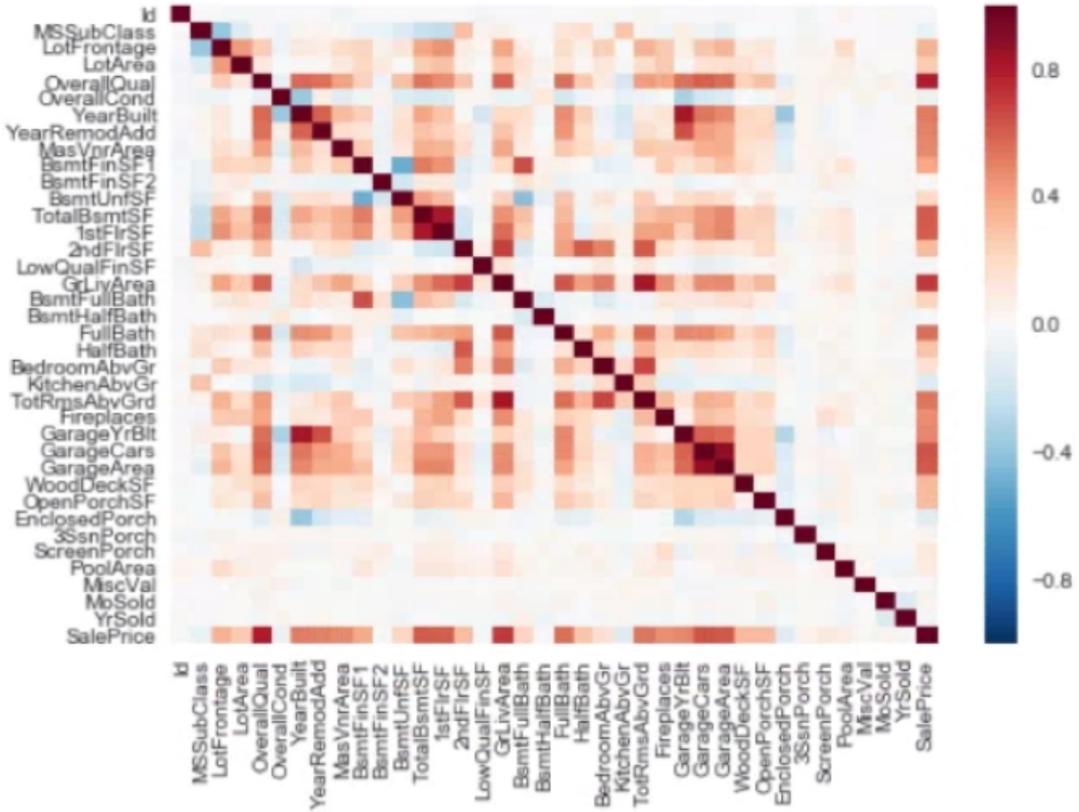
特徵選擇概念

- 特徵需要適當的增加與減少，以提升精確度並減少計算時間
  - 增加特徵：特徵組合 (Day 26) ，群聚編碼 (Day 27)
  - 減少特徵：特徵選擇 (Day 28)
- 特徵選擇有三大類方法
  - **過濾法 (Filter)**：選定統計數值與設定門檻，刪除低於門檻的特徵
  - **包裝法 (Wrapper)**：根據目標函數，逐步加入特徵或刪除特徵
  - **嵌入法 (Embedded)**：使用機器學習模型，根據擬合後的係數，刪除係數低於門檻的特徵
- 本日內容將會介紹三種較常用的特徵選擇法
  - 過濾法：相關係數過濾法
  - 嵌入法：L1(Lasso)嵌入法，GDBT(梯度提升樹)嵌入法

相關係數過濾法 (1 / 2)

下圖是探索資料分析(EDA)常見的相關性熱圖 (語法詳見於今日範例)

想想看我們該如何從圖上決定，該刪除哪些特徵?

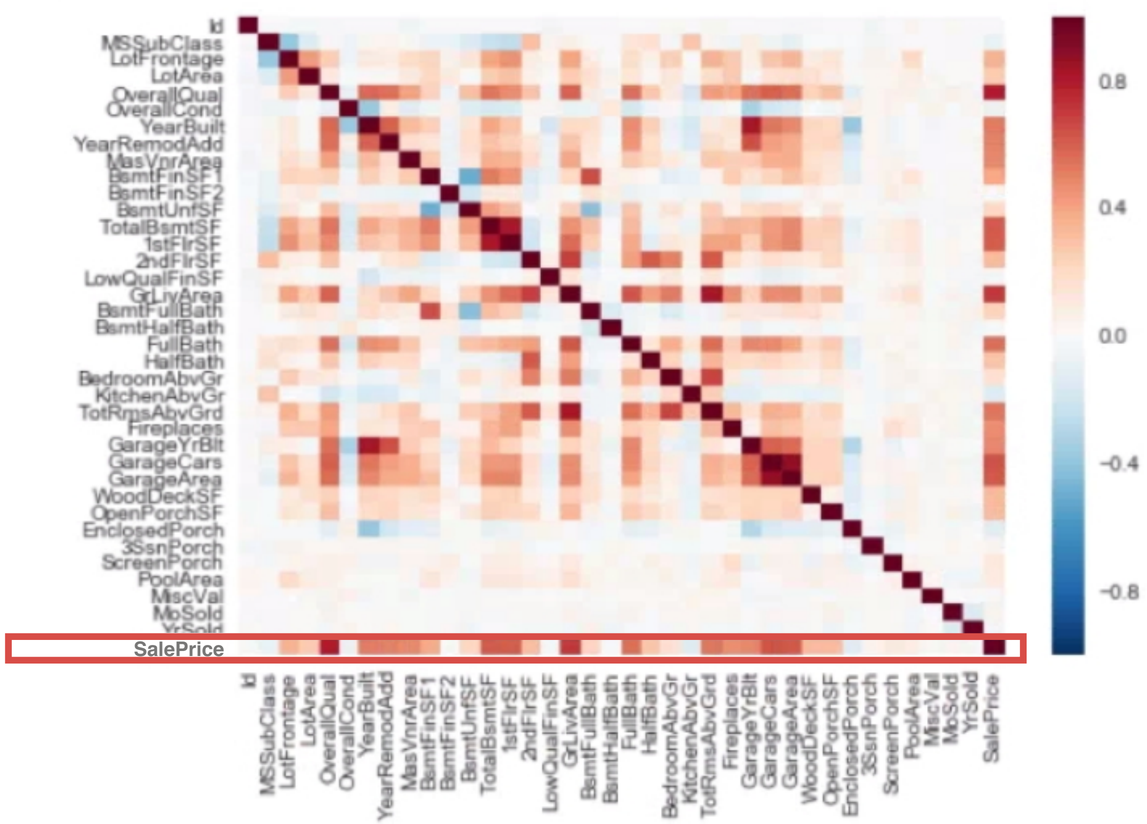


相關係數過濾法 (2 / 2)

- 找到目標值 (房價預估目標為 SalePrice)之後，觀察其他特徵與目標值相關係數 (紅框處)

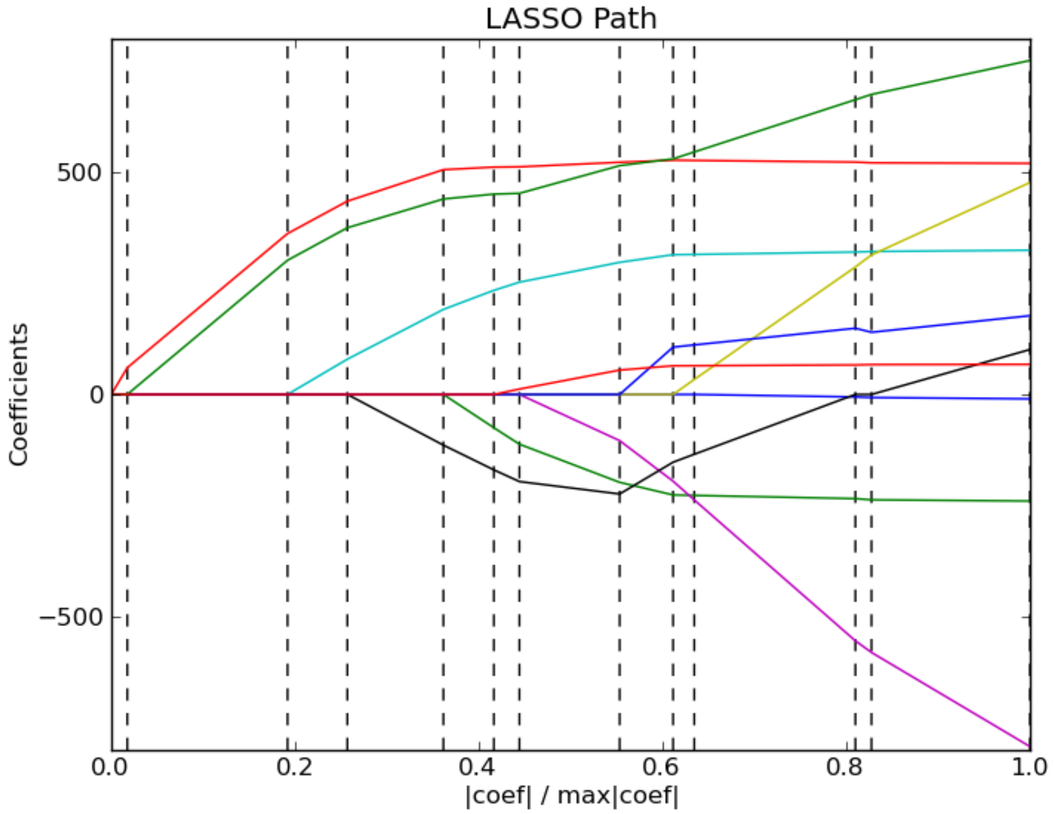
- 預設顏色越紅表示越正相關，越藍越負相關

- 因此要刪除紅框中顏色較淺的特徵：訂出相關係數門檻值，特徵相關係數絕對值低於門檻者刪除



Lasso(L1) 嵌入法

因為使用 Lasso Regression 時，調整不同的正規化程度，就會自然使得一部分的特徵係數為 0 ，因此刪除的是係數為 0 的特徵，不須額外指定門檻，但**需調整正規化程度**



圖源來源 : sklearn網站

GDBT (梯度提升樹) 嵌入法

- 使用梯度提升樹擬合後，以特徵在節點出現的頻率當作特徵重要性，以此刪除重要性低於門檻的特徵，這種作法也稱為 GDBT 嵌入法
- 由於特徵重要性不只可以刪除特徵，也是增加特徵的關鍵參考，因此我們會在 Day29 特別用一天為各位詳述特徵重要性

	計算時間	共線性	特徵穩定性
相關係數過濾法	快速	無法排除	穩定
Lasso 嵌入法	快速	能排除	不穩定
GDBT 嵌入法	較慢	能排除	穩定

上面是提到的三種特徵選取比較，看似各有長處，但是近年來GDBT的改良版本：Xgboost...等幾種算法，大幅改良了計算時間，因此成為了特徵選擇的主流

解題時間
It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題