Lab1報告

組員：108552005 高儀津(Luca\_Kao) , 108552017 申自強(Arthur.Shen)

1. 前言 :

在本次的作業House Prices: Advanced Regression Techniques 主要會使用多個房地產參數(約78個) , 對房地產的進行房屋SalePrice的預測。所以進行預測訓練前會對資料進行過濾或前處理。本次會嘗試使用不同方法像是Keras , lightgbm , xgboost 做訓練和分析。

1. 資料預處理(資料視覺化->數據資料清理->特徵工程) :

2.1 資料視覺化 — 使用SeaBorn套件是以 matplotlib 為基礎建構的高階繪圖套件，讓使用者更加輕鬆地建立圖表將資料視覺化。

2.2 數據資料清理 —

2.2.1 : OutLiers (離群值處理) - 去掉跟房產資料的分佈沒有正相關的值。

2.3.2 : Histogram & Normal probability plot(直方圖和常態機率圖分佈) -

透過此兩種圖觀測數據資料偏度是否有接近常態分佈(越接近對角線越好)。

2.3 特徵工程 —

2.3.1 : missing values (空缺值處理) - 檢查空值資料。

2.3.3 : Heatmap(熱圖) - 利用此圖檢查特徵數據相關度(矩陣)。

2.3.4 : 缺失值填充 - 根據每個特徵的現實含義, 填充合適的值。

2.3.5 : 數值轉類別 - 將數值類型轉化為字符串類型。

2.3.6 : Label encoding(類別轉數值) - 標籤編碼對不連續的數字或者文本進行編號。

2.3.7 : 新增特徵(TotalSF) - TotalBsmtSF + 1stFlrSF + 2ndFlrSF。

2.3.8 : 計算特徵偏度 - 對偏態分佈的數據進行標準化處理，使其更加服從正態分佈。

2.3.9 : 熱編碼 - 用來解決類別型數據的離散值問題(類別轉為無序 Dummy variables)。

2.3.10 : 特徵選擇 - 利用Lasso。

1. 模型選擇和評估 :

3.1 Lasso - (實作參考 : Luca\_lab1.ipynb)

3.2 ElasticNet - (實作參考 : Luca\_lab1.ipynb ,Arthur.Shen\_lab1.ipynb)

3.3 KernelRidge - (實作參考 : Luca\_lab1.ipynb)

3.4 Gradient Boosting - (實作參考 : Luca\_lab1.ipynb)

3.5 Xgboost - (實作參考 : Luca\_lab1.ipynb)

3.6 LGBM - (實作參考 : Luca\_lab1.ipynb)

3.7 Deep Neural Networks - (實作參考 : lab1\_DNN.ipynb)

3.8 模型疊加 - (實作參考 : Luca\_lab1.ipynb)

1. 討論 :

* 本次最終上傳到Kaggle Rank成績為0.12135(採用模型疊加方式)，此次的評估模型的準確性主要的差異為處理異常空缺值，偏度校正，特徵組合和選擇。
* 使用單一模型學習預測的效果並沒有多模型學習預測的效果好（可能是因為我們融合了幾個負均方差誤差比較小的模型來優化學習效果）。
* 對於Deep Neural Networks在這個數據集的訓練和處理，在前處理與特徵工程與其他模型處理相同的狀況下，效果並沒有上述的模型疊加方式好，可能是因為我們本次訓練使用的數據集非常小，因為一般來說當DNN擁有豐富的數據集訓練預測時，神經網絡的表現最佳。(目前最佳上傳至Kaggle的成績為0.17525, batch\_size = 16 , epochs = 1000 , 當epochs >400 就無法將數據收斂了，還需要其他方法改進實驗)。

5. 參考資料:

<https://github.com/lars10192002/Introduction_to_Deep_Learning/blob/master/Lab1/Luca_Notes/note.md>