DataMining Pa1 Report

一、實驗方法

（一） 資料格式整理

i. T/F 資料值：將文字屬性資料以 1/0 取代

ii. 為了方便進行下一步的資料挑選，以及觀察 X 與 Y 的關聯性，將 Y_train 資料與 X_train 合併

（二） 資料清理

i. 首先從月租金的資料分佈圖發現，月租金有資料有一個明顯的極值，因此先刪除此項資料

ii. 從資料分佈中觀察到，台北市的租金明顯高於其他縣市，而新北市位區第二，因此將縣市做一個「縣市分類」的 dummy variable

1. 台北市為第一類
2. 新北市為第二類
3. 其餘則分為第三類
4. 經過此處理後，縣市與月租金之間的相關係數從 0.067 上升到 -0.466

iii. 再來是觀察到衛浴數、陽台數過高的資料對於月租金並無影響，故將其資料值取代為眾數。

iv. 另外觀察到月管理費大於月租金，不合邏輯，可能是錯誤資料，因此將其刪除。

v. 樓層數部分依台灣大部分的建築情形以及建築法規，刪除掉大於 40 以及小於-6 的樓層位置

（三） 特徵選擇

i. 在特徵選擇的部分使用了 RandomForestRegressor，先將所有的特徵值與月租金做第一次擬合，並將特徵重要性進行排序。

ii. 挑選出前一半的屬性做訓練

（四） 訓練模型

i. Random Forest

1. 使用 Sklearn 的套件，樹為 300 棵，深度 40 跟沒有深度限制之結果差不多

ii. Adaboost

1. 同樣使用 Sklearn 的套件，基於樹的棵數同樣為 300

二、實驗結果與分析

（一） 在 randomforest 設定完全相同的情況下，不對資料進行任何處理（只修改資料型態）的分數，與最後做出的結果相差並不大。推測是沒有找到合適的 dummy variable 將部分資料做進一步的調整，另外特徵選擇的數量可能也是影響結果的原因之一。

（二） 在嘗試特徵選擇時，有嘗試使用 PCA，但反而使結果變差。

（三） 因 RandomForest 不受到屬性的 Scale 影響，所以進行標準化後區別不大。

三、參考資料

（一） https://colab.research.google.com/github/AI-FREE-Team/Data-Analytics-in-Practice-Titanic/blob/master/Data%20Analytics%20in%20Practice%20-%20Titanic%20Survival%20Prediction.ipynb#scrollTo=hzGr6VAApIUB