Practice of Social Media Analytics HW1 – Link Prediction

M11207509 王佑強

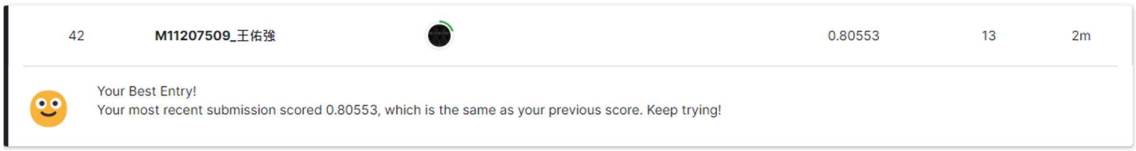
程式執行方法:

1. 使用 jupyter(此檔案為 ipynb)
2. 安裝套件 numpy、pandas、sklearn、 xgboost、networkx
3. 可分別執行 SVM、RandomForest、xgboost 下的 code block

- 分數最高為: xgboost

演算法流程:

1. 將訓練資料及測試資料讀進來, drop 掉不需要之欄位()
2. 將訓練資料分為訓練集以及驗證集
4. 測試資料集只留下特徵: node1、node2
5. 分別將訓練資料丟入 SVM、RandomForest 以及 Xgboost 去訓練
6. 用 networkx 建立有向圖網路, 紀錄特徵: s**uccessor_count, predecessor_count**
6. 使用測試資料去預測最後結果
7. 將預測結果與 sample_submit.csv 合併產生完整可交至 Kaggle 的檔案



心得:

1. 一開始 SVM 時準確度只有 0.49, 使用 GridSearchCV 尋找最佳參數, 調整為 C=0.1, gamma='scale', kernel='poly' 0.57

	precision	recall	f1-score	support
0	0.60	0.83	0.70	5760
1	0.41	0.17	0.24	3840
accuracy			0.57	9600
macro avg	0.51	0.50	0.47	9600
weighted avg	0.52	0.57	0.52	9600

2. 而 RandomForest 是一種集成式學習(Ensemble Learning)的方法，由多棵決策樹(decision tree)組成從資料集中挑 n 筆資料中隨機挑選 k 個做為樣本重複 m 次，產生 m 個決策樹分類，以多數決制進行預測 結果為: 0.61

3. 第三種使用 xgboost, 一開始 score 為 0.615, 用 networkx 建立有向圖網路，再新增兩個特徵分別為 **successor_count**: 表示有多少箭頭是從這個節點指出去的。 例如：如果有一個節點 A，從 A 指向 B 和 C 的話，則 A 的後繼者數量為 2。 **predecessor_count**:如果有一個節點 B，並且 A 和 C 兩個節點都有指向 B 的箭頭，則 B 的前驅者數量為 2。**可以幫助理解節點在圖中的角色與影響力**，例如在社交網絡、網站結構或任何其他有方向性的關係圖中。在預測任務中，這些特徵可能會提供關於節點之間關係強度或可能性的重要信息。最後結果為 0.80553