機器學習導論

Homework #3
Due 2023 Nov 6 11:00PM

(一) 針對員工離職率(left)進行離職與否的預測

資料檔案：*HW2_hr-analytics_train.csv, HW2_hr-analytics_test.csv*

作業要求：

1. 讀進訓練資料 *HW2_hr-analytics_train.csv*，判斷出那些數據格式不是數字，或是有缺失值。
2. 將非數字類型的資料進行必要的編碼。
3. 若有缺失值請填補。
4. 建立 **Decision Tree** 模型並進行訓練。請呈現訓練後模型預測的混淆矩陣。
5. 試著找出最重要的前兩種特徵，請說明你如何找出特徵之重要性。
6. 請利用訓練後的模型預測測試資料 *HW2_hr-analytics_test.csv* 的離職情況，並將結果存成 *HW2_hr-analytics_test_sol.csv*，儲存格式如下範例。該結果的準確率將佔此一題分數的 ***35%***。

	A	B
1	left	
2	1	
3	0	
4	1	
5	1	

7. 請與前次 Logistic Regression 的預測準確率進行比較。請探討那個模型比較適合，其可能原因為何?

(二) 針對信用卡交易資料，預測是否為詐騙的交易（class==1）

資料檔案：*HW3_creditcard.csv*

作業要求：

1. 讀入資料、切割資料（測試集佔 30%，訓練集佔 70%）
2. 利用 Decision tree 進行預測，計算出 Accuracy, Recall, Precision, F1-Score 及 AUROC。
3. 統計 class==0 及 class==1 的資料筆數，看是否類別間資料數量是否有很不平衡的現象。
4. 為了要提高 recall 的數值，請：
   - 改變 Decision tree 模型中類別權重或訓練權重，計算新的結果，與之前結果比較。
   - 利用 imbalanced-learn 套件中 SMOTE 的方法來增量資料，計算新的結果，與之前結果比較。
5. 改利用 xgboost 模型重新做 step 2。

繳交說明：請繳交 jupyter notebook 之檔案。若有討論部分也利用 jupyter notebook 說明。