4107056006 資工三 游庭瑋

HW#5 Logistic regression Interview

1. 作業目標：  
   利用pandas及sklearn模組，預測給定數據的y預測值
2. 資料集：
   1. interviewTrain.csv 內有：
      1. height
      2. weight
      3. handson （此處資料有誤植，應為handsome）
      4. rich（此處在單字後方多了一個空白）
      5. experience
      6. y

此資料為用來訓練的資料集

* 1. interviewTestEven.csv 內有：
     1. height
     2. weight
     3. handson （此處資料有誤，應為handsome）
     4. rich（此處在單字後方多了一個空白）
     5. experience
     6. y

此資料為用來測試用

* 1. interviewTrainStandard.xlsx 內有：
     1. height
     2. weight
     3. handson （此處資料有誤，應為handsome）
     4. rich（此處在單字後方多了一個空白）
     5. experience
     6. y

其中y值為公式完成的，需要符合以下條件：

For each row[i]

if row[i].height > 40 then count += 1

if row[i].weight > 30 then count += 1

if row[i].handsome > 60 then count += 1

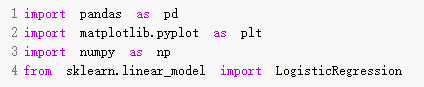
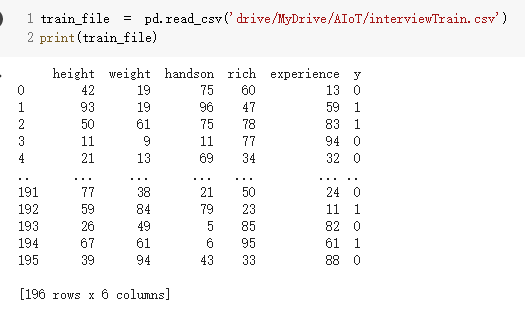
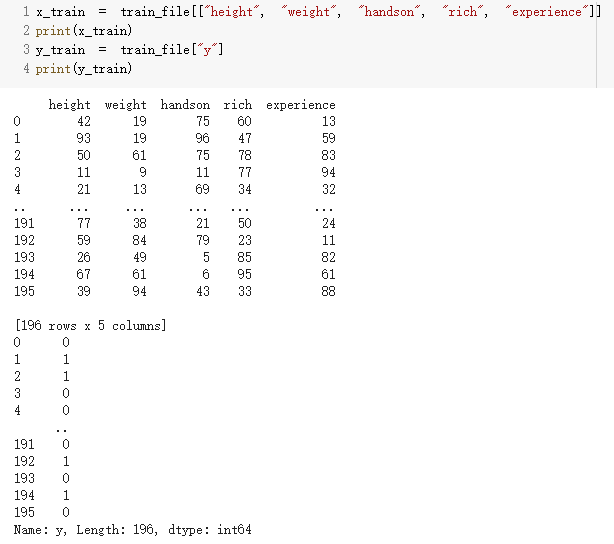
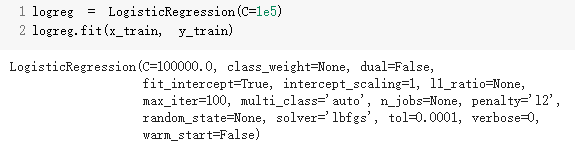
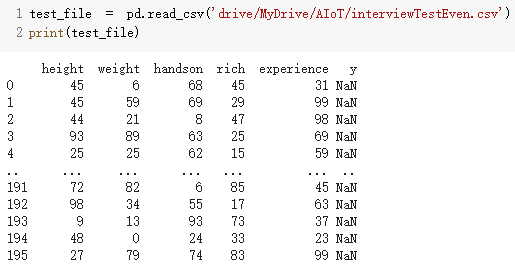
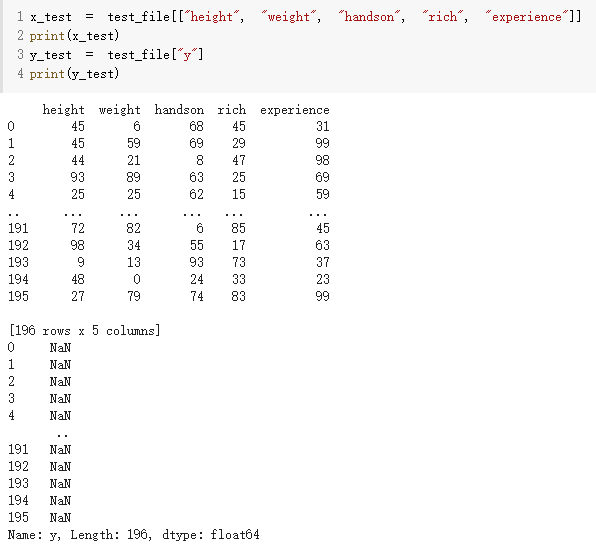
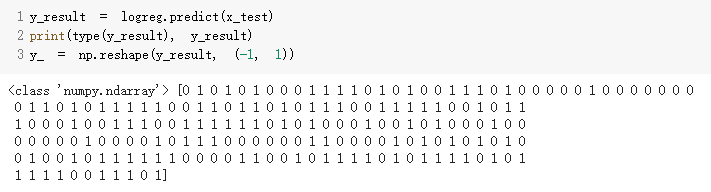
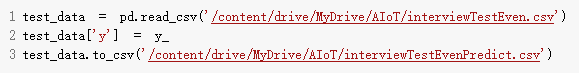
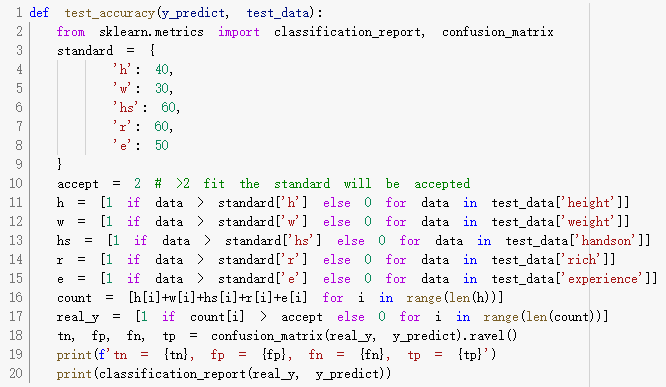
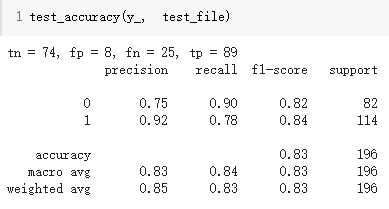
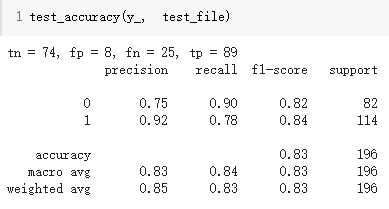
if row[i].rich > 60 then count += 1

if row[i].experience > 50 then count += 1

if count >= 2 then y := 1

else y := 0

此資料用來驗證計算的正確性

1. 操作環境：  
   本次作業使用colab來實做，以下資料有誤的部份皆以原資料為主
2. 操作步驟：
   1. 先引用pandas、sklearn等套件  
      
   2. 載入訓練資料集並確認  
      
   3. 將訓練資料的x, y分開  
      
   4. 針對上述資料的x, y做羅吉斯回歸的訓練，其中給定的C值代表正規化強度的倒數，越大代表強度越低  
      
   5. 接著將測試資料匯入  
      
   6. 切分測試資料的x, y  
      
   7. 利用訓練好的資料，預測測試資料集  
      
   8. 重新載入測試資料集，並寫入預測y值  
      
   9. 根據interviewTrainStandard.xlsx的資料，可以用下述函數來驗證預測的正確性  
        
      其中3～17行是根據interviewTrainStandard.xlsx的規則作為基礎，來產生對應的正確y值；18～19行利用sklearn.metrics套件中的confusion\_matrix來生成混淆矩陣；20行利用sklearn.metrics套件中的classification\_report來產生precision, recall rate等資料，執行函數結果如圖：  
      
3. 結果分析：  
   

透過上表可以看出用這筆訓練資料來預測的結果，其中，accuracy可定義為所有情況中正確判斷的比例、precision為所有判斷為真的情況下真的為真的比例、recall為所有真的情況下有多少被判斷出來。可以看到在1的precision達到0.92，代表model在判斷是1的情形下確認真的是1的比例很高，相較之下0的precision就沒那麼高了；但在recall上，反而是0比較高，代表著model在判斷正確的情況下確認是0的比例較高。

1. 結論：  
   這次作業透過python sklearn的logistic regression來在沒有給定條件的情況下直接預測y值，有點類似用答案推題目的感覺。也有發現在logistic regression上還是有可能出現判斷準度較低的情形，如果提高訓練資料的數量，應該可以更好的去預測數值。