Machine Learning on UCI Adult data Set

by

紀明妤

摘要

這份報告是為了呈現如何使用機器學習，使用邏輯回歸的特徵工程來解決問題。

介紹（研究背景及研究目的）

近幾年因科技進步及資料來源的快速增長，機器學習領域已經發展成我們不僅僅能從資料庫中提取我們想讀取的資訊，還能預測未來的趨勢，而這是在 10 年前數據還沒有這麼蓬勃發展的時代很難達成的。

機器學習的流程大致都包含以下內容：轉換資料成可讀取的型式、資料清洗、以及針對資料做出相對應結論。

兩個機器學習的主要子領域包含監督式學習和非監督式學習。前者主要有決策樹(Decision trees), 單純貝氏分類器(Naïve Bayes classifier), KNN 演算法(K-Nearest Neighbor algorithm), 邏輯迴歸(Logistic Regression)及支援向量機(Support Vector Machine)。

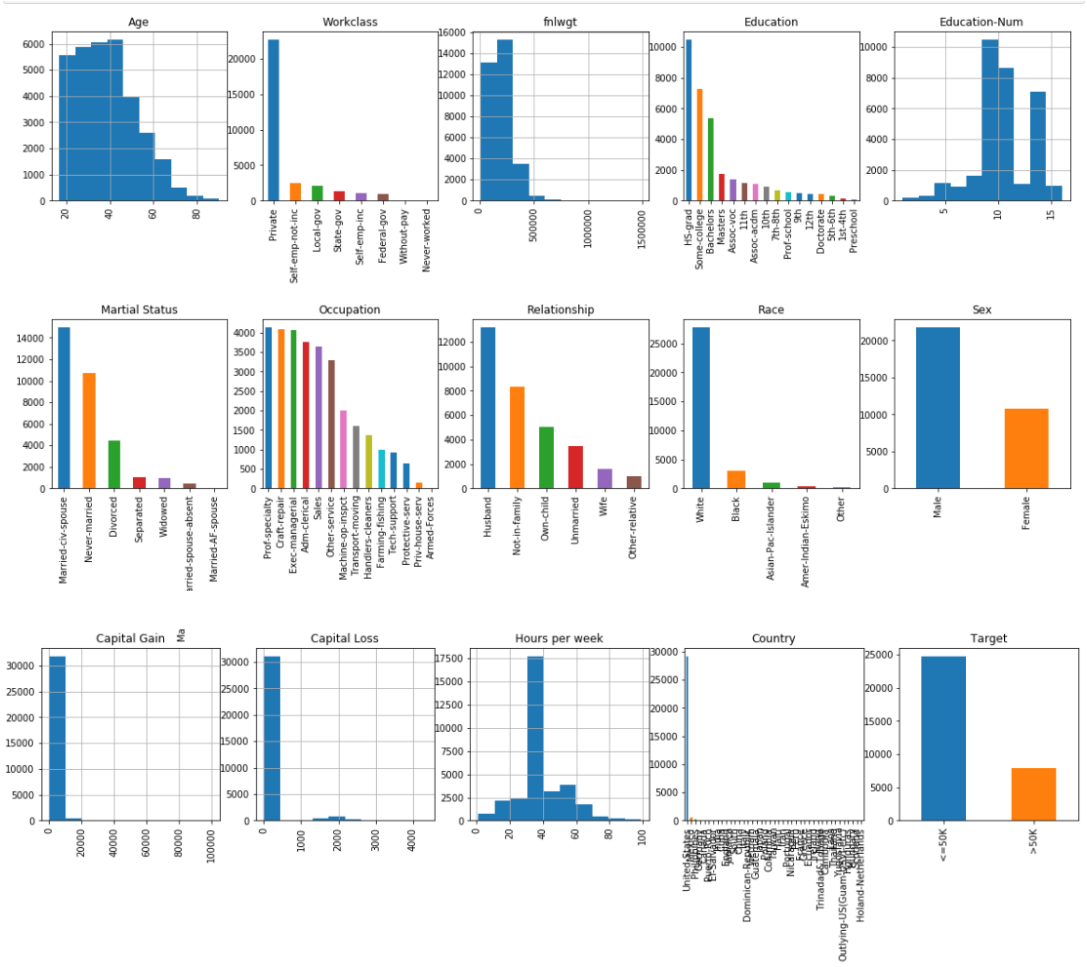
在這份報告中，研究目的是使用邏輯回歸演算法，來提升預測準確值。

資料集介紹(含資料特徵)及資料集來源

此分析報告的資料集源自於 UC Irvine University Machine Learning Repository 的“Adult data set”。這個資料集屬於二元分類問題-針對 48842 人的數據集，給定教育、性別等資料，共有 14 個屬性，其中可能相依或獨立，建構機器訓練模型(邏輯回歸)，並預測每一人年收入是否超過 5 萬美元。

資料預處理

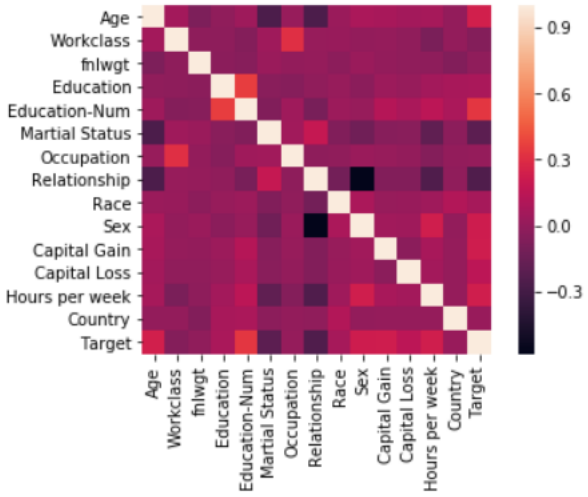
導入數據後，我們繪製每個特徵的分布，因為這能讓我們能更好的理解數據。



從數據特徵分布我們可以得知，數據主要集中於美國，且是美國白人男性。美國的數據樣本有 89，而來自墨西哥的數據樣本只有不到 2%。

```
United-States    0.895857
Mexico           0.019748
Philippines      0.006081
Germany          0.004207
Canada           0.003716
Name: Country, dtype: float64
```

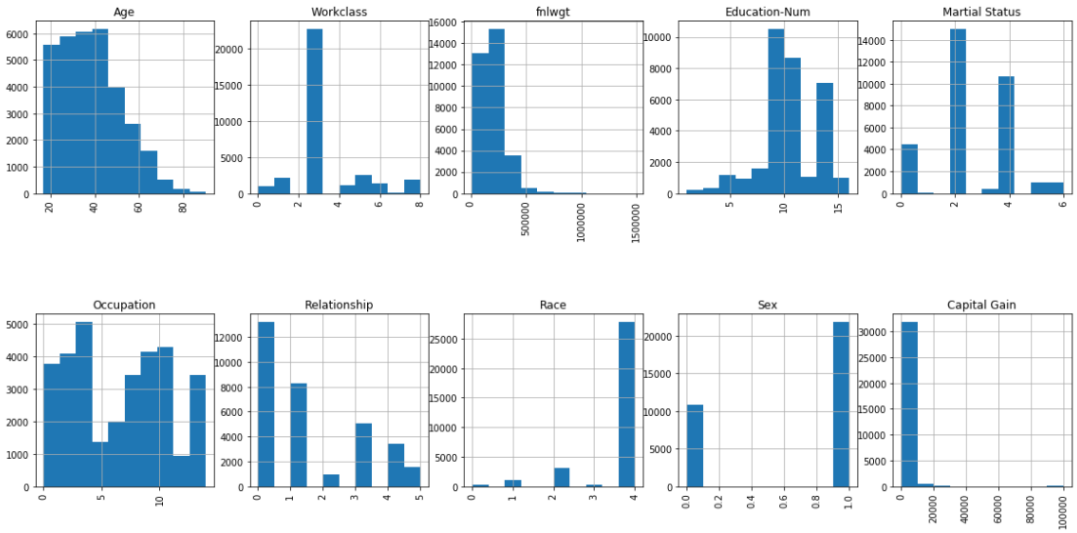
接著我們探討不同特徵之間的相關係數。通常我們不樂見不同特徵之間存在很多相關性，因此我們需要使用 scikit-learn package，發現 Education 與 Education-Num 具高度相關性。這兩列代表相同功能，但編碼分別為 string 和 number。我們選擇刪除 string 的 education 欄位，因為 Education-Num 有重要的排序特質-數字越高，一個人的教育程度越高。這是一個機器學習演算法能運用的重要資訊，使刪除 education 後的資料看似呈現負相關。

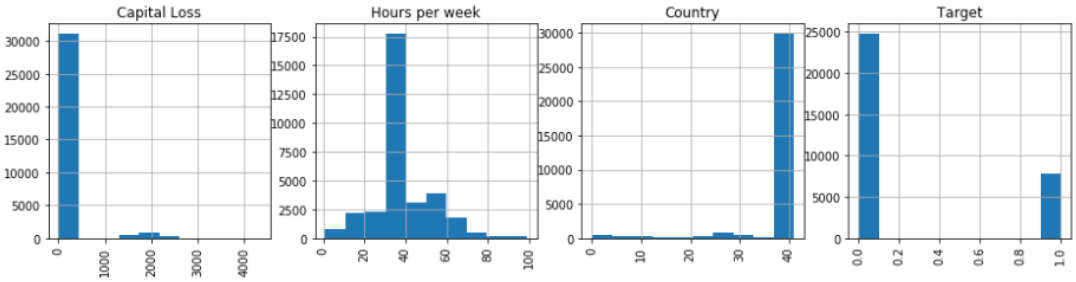


機器學習或深度學習方法（使用何種方法）

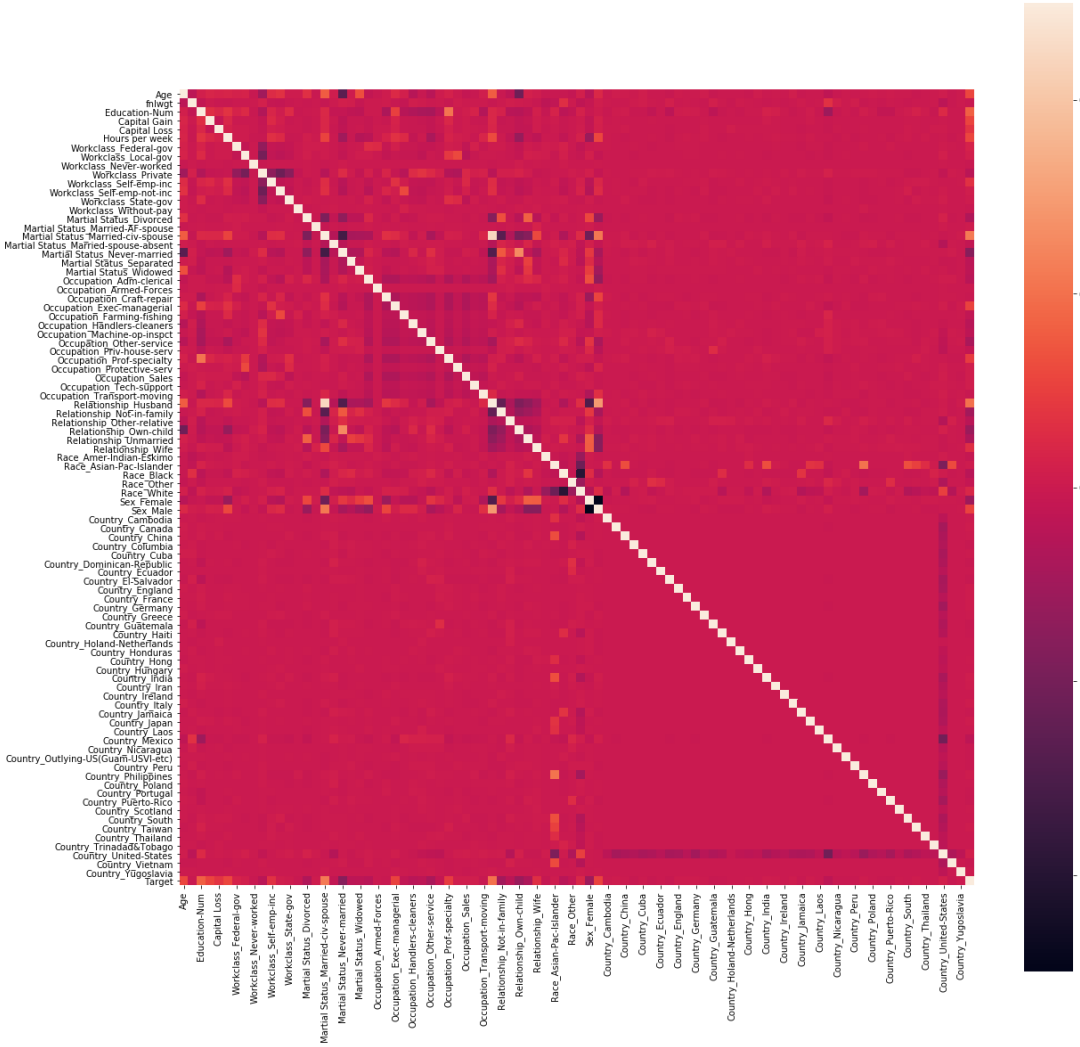
為了研究擁有的數據集，我們嘗試建立一個分類器，給定數據來預測既定人士的收入。

首先，建立一個分類器時，我們需要將特徵從 string 轉化成 number，並將資料分類成訓練集和測試集，避免過度擬合。

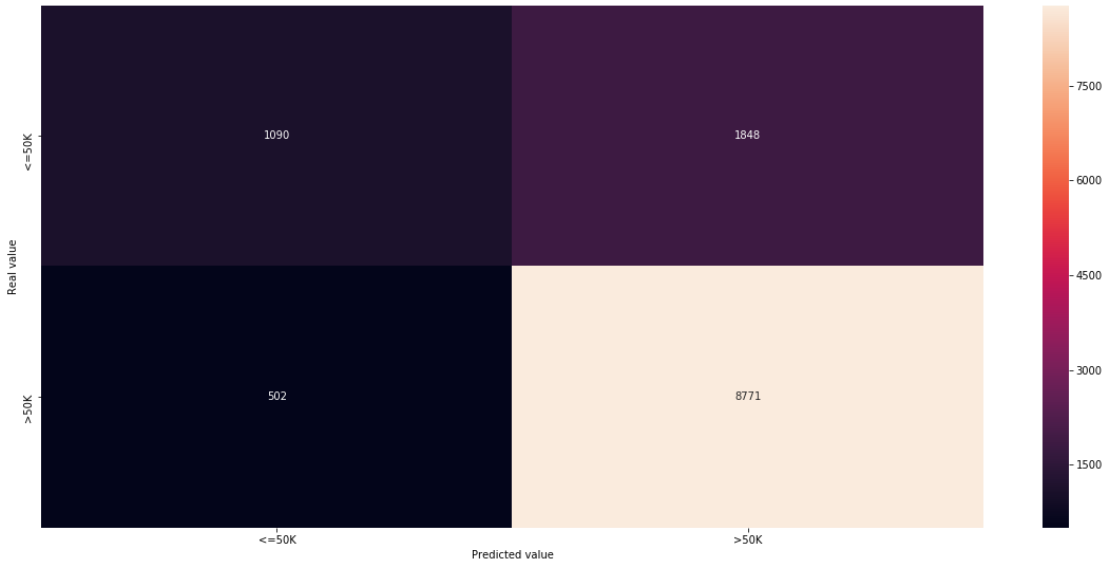




最後，為了不讓 logistic regression 認為我們以數字 1-5 來 encode data"relationship"有不同的權重，因此我們使用 dummy variables 來讓權重相等



並繪圖標示出正確預測 income<=50K 或>50K 的確切數量。



研究結果及結論（含模型評估與改善）

因發現 logistic regression 在辨識資料時會依據我們的編碼來給予不同的權重，然而有些資料如 ”Marital Status” 將數值分類從 0 到 6 ，並沒有像 education_num 的數值一樣具有意義，故使用 dummy variables 來改善此缺點並提高模型預測值。

參考文獻

http://mlr.cs.umass.edu/ml/datasets/Adult