巨量資料分析期末專題

人口普查資料分析

資四甲 ACS105114 吳佳恩

資四甲 ACS105138 陳奕安

資四甲 ACS105144 洪英祐

目錄

[**一、摘要**](#_3drqziub9re8)1

[**二、使用資料集**](#_6taulz2mj91a)1

[**三、特徵分析**](#_dav6xqppfx25)2

[**四、深度學習模型**](#_kkbfkux1oj2h)5

[**五、訓練模型分析結果**](#_e5rbogagnene)5

[**六、結論**](#_c6gcqkac2un)6

[**七、參考資料**](#_xjbezlwq6q07)7

# 一、摘要

我們的專題使用深度學習的方式分析人口普查資料集，目的是透過提供的3萬多筆資料的14個特徵值訓練一個分類器來分辨個人年所得是否會高於5萬美金。

在特徵選取上我們使用了資料統計和線性迴歸的方式判斷特徵的重要性，並使用了R的Boruta package確定我們的評估結果可靠，最終透過Caret package的兩種深度學習Model實際訓練分類，證明了我們的特徵選擇確實找出了比較高相關性的重要特徵，以較低的訓練成本達到更高的效益。

# 二、使用資料集

我們使用的資料集是UCI的Census Income Data Set，它記錄了西元1994年美國的人口收入普查資料，總共有32561筆訓練資料及16281筆測試資料。

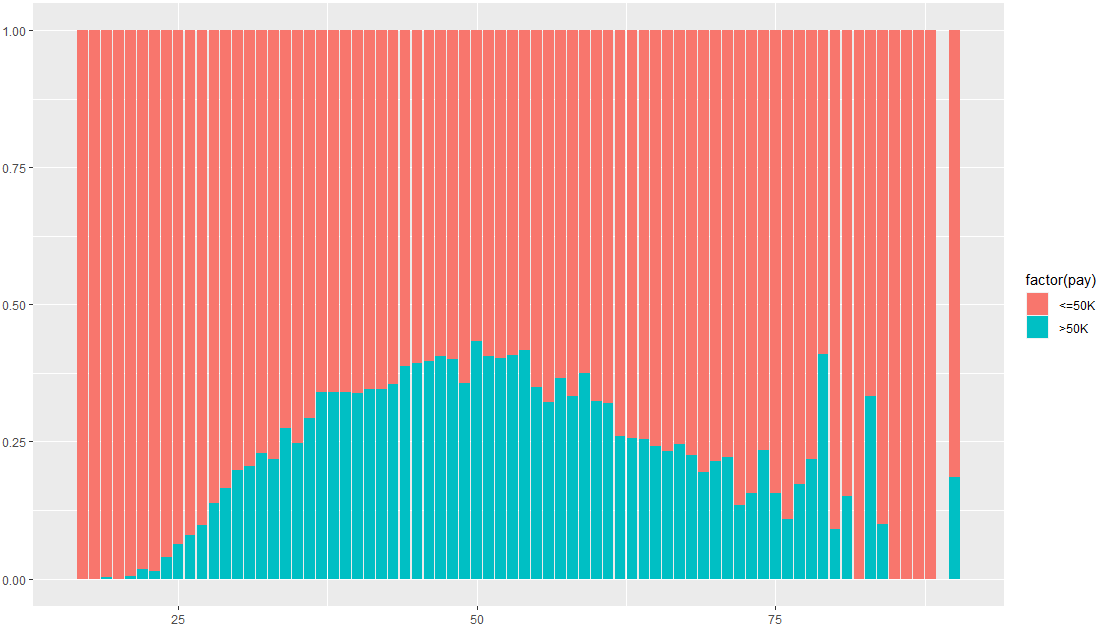
這個資料集每個記錄都包含了15種個人資訊，分別是我們要輸入學習的14個特徵值，以及我們要判斷的收入標籤。

表一、資料集特徵

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | age | 年齡 | 9 | race | 種族 |
| 2 | workclass | 工作類別 | 10 | sex | 性別 |
| 3 | fnlwgt | 序號 | 11 | capital-gain | 資本收益 |
| 4 | education | 教育程度 | 12 | capital-loss | 資本損失 |
| 5 | education-num | 受教育時長 | 13 | hours-per-week | 每周工時 |
| 6 | marital-status | 婚姻狀況 | 14 | native-country | 國籍 |
| 7 | occupation | 職業 | 15 | pay | 年收入 |
| 8 | relationship | 家庭關係 |  |  |  |

# 三、特徵分析

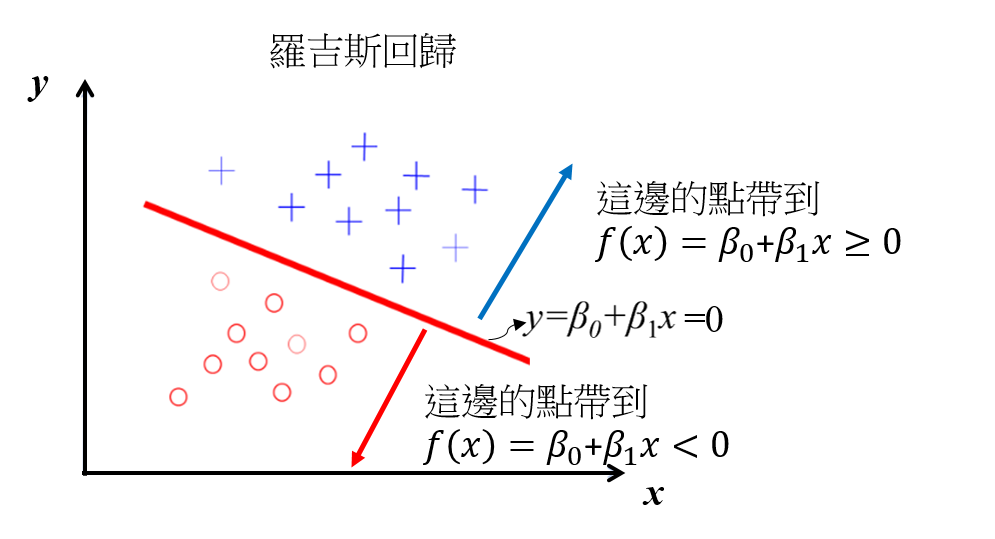
我們目標的判定主要是一個人的年收入是否超過50k，而判斷的依據總共有14種特徵值，而首先我們先預測哪些特徵值對目標判定的影響是較大的，我們首先選出了"每週工時"、"種族"、"年齡"、"受教育時長"等重要特徵來與目標特徵做比較，畫出了百分比堆疊直方圖來判斷這些特徵是否真的是重要特徵，得到了以下結論(以年齡為例)，



圖一、年齡與年收入堆疊圖

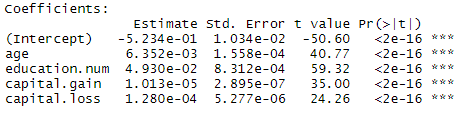
由圖可知年齡在進入中年時年收入大於50K的比例是明顯大於其他年齡層的，這邊可以推測在青年時期大部分的人可能還在進修或者事業剛起步可以解釋為何收入較低，而老年也就代表進入退休生活，這也間接證明了年齡跟年收入之間是有一定關係的。

為了證明我們的重要特徵的選擇是正確的，我們接著再使用了羅吉斯迴歸(Logistic regression)證明我們選出的特徵與目標的關係，羅吉斯迴歸主要是用用線性迴歸的輸出來判斷這個資料屬不屬於target，簡單來說就是將點(數據)帶進去迴歸線，迴歸線輸出值若是>=0是一類(target)，值<0是另一類(non-target)

****

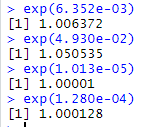
圖二、羅吉斯迴歸表示圖

以下是以年收入當目標與年齡做羅吉斯迴歸的執行結果



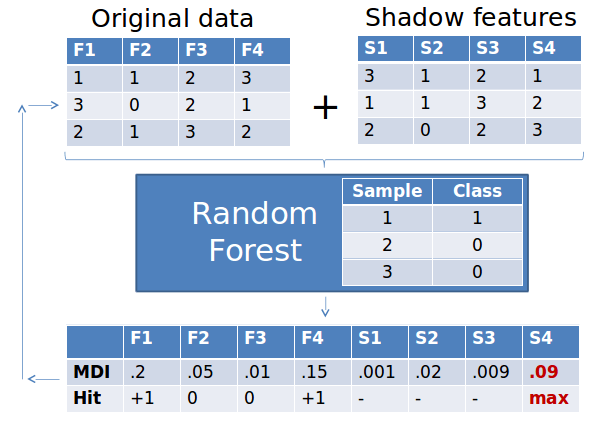
圖三、羅吉斯迴歸結果

在將age迴歸係數進行指數運算可以發現結果是趨近於1，也就代表與目標(年收入)的關係是十分密切的。



圖四、age係數指數運算值

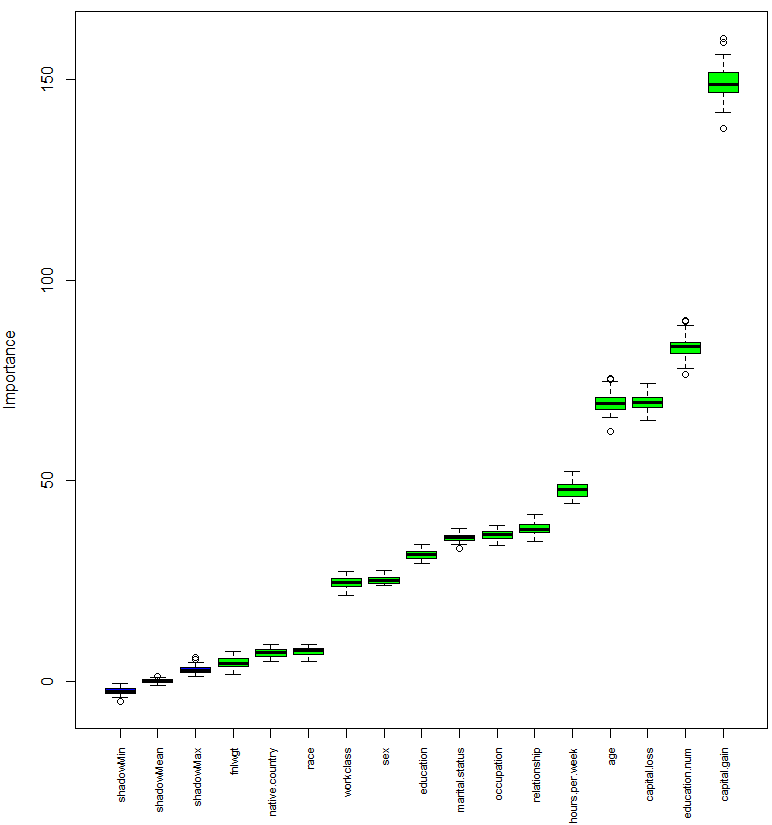
接著我們再使用一種叫Boruta的演算法，Boruta這個方法是基於random forest classification產生的一種演算法，它的目標是找出在高維的訓練資料中對辨識度影響較大的特徵去做特徵選取，藉此來提高模型訓練效果。



圖五、Boruta Importance計算

這種演算法將original data進行shuffle後產生出shadow features，再將original data與shadow features拼接作為特徵矩陣進行訓練，然後根據original data和shadow features去計算分數比對產生出各feature的importance得分，找出對訓練較為重要的特徵集合。

如圖六所示，從計算結果中可以發現，影響最大的前幾項與我們最初所推估出的重要特徵基本符合。



圖六、14項特徵的Importance值

# 四、深度學習模型

我們使用了R語言的Caret package來做我們的機器學習神經網路，我們選擇了兩種神經網路，線性判別分析(Linear Discriminant Analysis, LDA)和CART(Classification and Regression Trees)。

LDA是一種線性分類的方式，它會把空間中的資料根據特徵值做線性區分。而CART是一種決策樹的分類方法，可以用於分類(離散類型)或迴歸(實數)預測模型問題，它會尋找目標屬性與目標值之間的映射關係來做分類辨識。

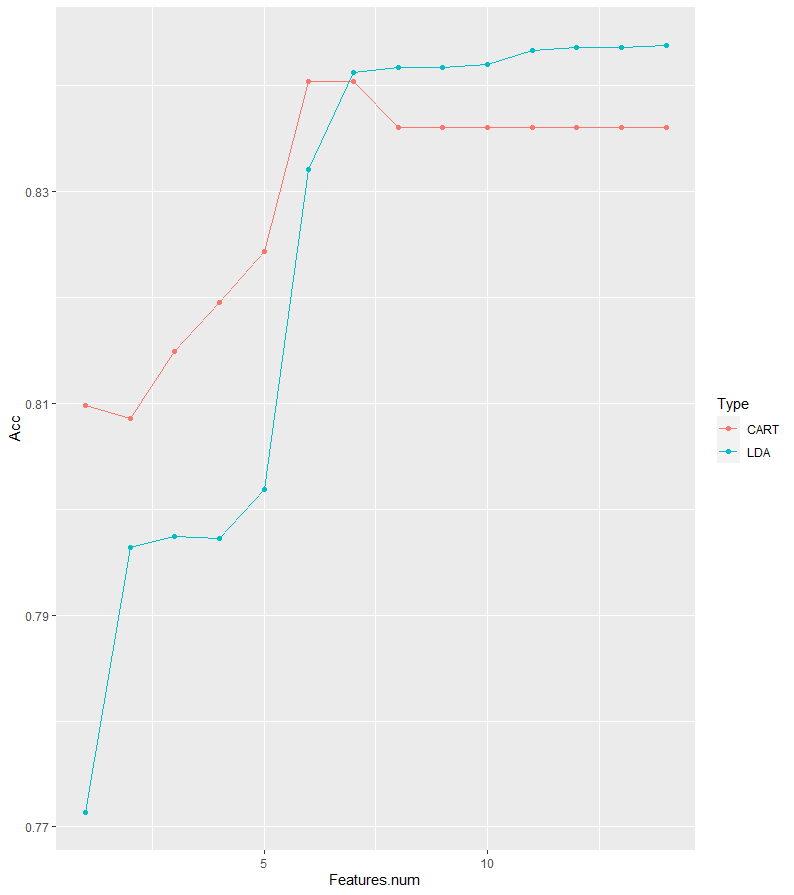
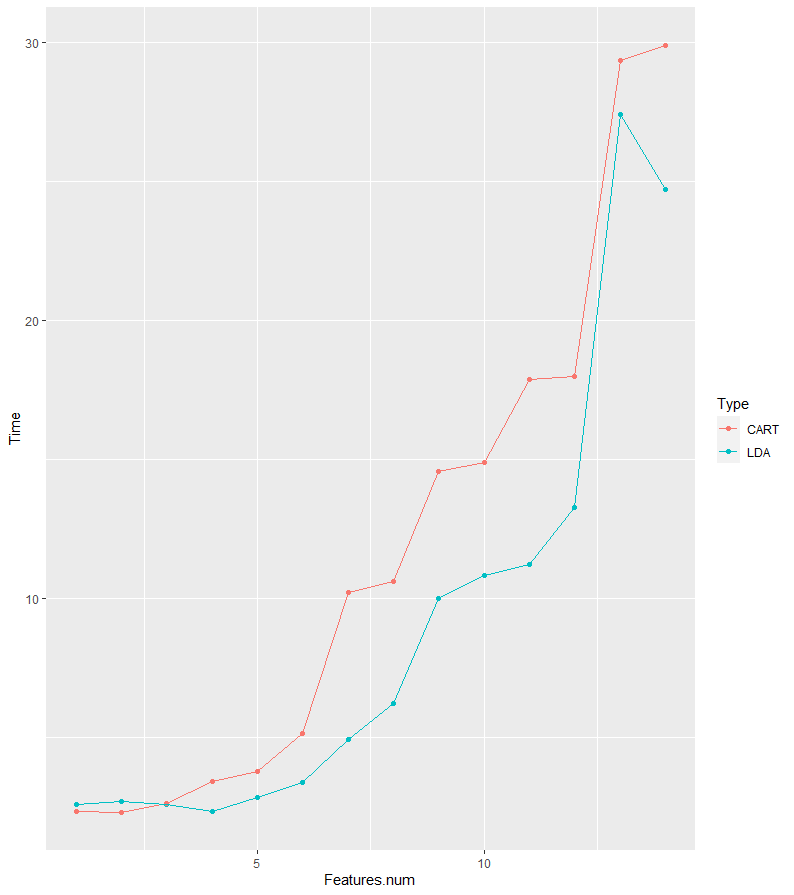
# 五、訓練模型分析結果

比較一下三個模型的訓練結果可以發現，雖然辨識準確度相差沒有很大，但是在特徵數相同的情況下使用高關聯性資料確實會有比較好的辨識效果。

而使用了所有特徵訓練的LDA模型準確度雖然略高於使用九個高相關特徵的LDA模型，但這差距並不是很大，在CART的模型中甚至達到了相同的數值。因此我們可以得知在這個資料集中每個特徵都多少對辨識有幫助，但低相關資料的影響十分微小，而扣除這些低相關特徵能讓我們節省1/3的運算量。

表二、模型辨識準確度

|  |  |  |
| --- | --- | --- |
|  | LDA模型ACC | CART模型ACC |
| 前九項特徵 | 83.18% | 82.72% |
| 九項高關聯性特徵 | 84.17% | 83.61% |
| 所有特徵 | 84.38% | 83.61% |



圖七、特徵數量與訓練時間、準確度關係圖

# 六、結論

根據最終的模型訓練結果可以證明特徵的選擇確實會影響到我們的分析模型準確度，使用的特徵數量也會對訓練時間造成高度影響。而我們選擇重要特徵的方法跟重要性評估計算方法也透過最終成果得到了驗證，證實我們的方法的確具有可靠性跟有效性，能在不過度增加運算成本的情況下達到較高效益。

# 七、參考資料

[1] UCI, ‘Census Income Data Set’

from:<https://archive.ics.uci.edu/ml/datasets/Census+Income>

[2] Jason Brownlee, ‘Your First Machine Learning Project in R Step-By-Step’

from:<https://machinelearningmastery.com/machine-learning-in-r-step-by-step/>

[3] Mehedi Hasan, ‘The 20 Best R Machine Learning Packages in 2020’

from:<https://www.ubuntupit.com/best-r-machine-learning-packages/>

[4] dattatrayshinde, ‘UCI Adult dataset / Census Income Classification‘

from:<https://github.com/dattatrayshinde/UCI-AdultDataset-ClassificationProblem>

[5] 數盟譯文, ‘數盟譯文--如何使用R語言的Boruta包進行特徵選擇’

from:<https://kknews.cc/zh-tw/tech/rxyblo.html>

[6] 田鑫, ‘特徵選擇利器: Boruta’

from:<https://zhuanlan.zhihu.com/p/33730850>

[7] Manish Pathak, ‘Feature Selection in R with the Boruta R Package’

from:<https://www.datacamp.com/community/tutorials/feature-selection-R-boruta>

[8] Wiki, ‘決策樹’

from:<https://zh.wikipedia.org/wiki/%E5%86%B3%E7%AD%96%E6%A0%91>

[9] Wiki, ‘線性判別分析’

from:<https://zh.wikipedia.org/wiki/%E7%B7%9A%E6%80%A7%E5%88%A4%E5%88%A5%E5%88%86%E6%9E%90>

[10] Tommy Huang, ‘羅吉斯迴歸(Logistic regression)’

from:<https://medium.com/@chih.sheng.huang821/%E6%A9%9F%E5%99%A8-%E7%B5%B1%E8%A8%88%E5%AD%B8%E7%BF%92-%E7%BE%85%E5%90%89%E6%96%AF%E5%9B%9E%E6%AD%B8-logistic-regression-aff7a830fb5d>

[11] glm function | R Documentation

from:<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm>