Homework 2 Report - Income Prediction

學號：r06521605
系級：土木所電輔組碩一
姓名：許舜翔

1.

以同樣的特徵組合來進行訓練，那以上傳 Kaggle 的結果來說，logistic regresstion 與 generative model 的準確率分別是 *0.85* 與 *0.80* 左右，是以後者的表現較差，推測可能的原因是，generative model 的實作方法包含對資料機率函數的假設，雖然能較快獲得結果且所需資料也相對較少，但這次拿到的資料集，其中兩者的比例明顯不均，*(<=50k)*與*(>50k)*的數量比大概是 *0.76:0.24*，這樣的分布也造成預先假設的機率函數過於高估*(<=50k)*的可能性，導致最後結果有較大的誤差。

2.

以下分點介紹我實作模型的方法，其中特徵選取原先擬採用相關係數的大小來進行挑選，但結果並不如預期，後來採用較土法煉鋼的方式，由於我認為數值資料的分佈與分類有關，故皆保留，去一一剔除其他類別特徵，來找到精度較佳的特徵組合。

- **特徵組合**：我選擇採用所有的數值資料，且剔除 *native_country, relationship, education, race* 這些特徵，再進行展開，最後得到數量為 39 的特徵組合，並利用 Keras 套件，設定兩層 Core (其激活函數第一層設為 relu，第二層設為 *sigmoid*)，第一層輸入及輸出維度皆為原先特徵數量，第二層則是輸出為一個值，即最後衡量應分為何種類別的依據。
- **訓練參數**：設定 *epochs* = 50, *batch_size* = 128, *validation_split* = 0.1。

3.

由於是採用 sigmoid 的函數來做特徵轉化，故若沒有事先做標準化的話，容易造成 exponential 的值 overflow，且會低估其他特徵的造成的影響，對 logistic function 的實作來說會導致 gradient 在該特徵的計算過大，修正的方向無法符合訓練資料的特性，導致模型會訓練不起來，以下為實作於 logistic regression，且固定其他訓練參數，僅考量有無標準化的精度比較表。

特徵組合中受標準化影響的特徵：*age, fnlwgt, education_num, capital_gain, capital_loss, hours_per_week*

	有 normalize	無 normalize
Accuracy (training set)	0.8458	0.7962
Accuracy (validating set, validation split = 0.1)	0.8434	0.7928

4.

以下列出在相同條件下（feature, learning rate, batch size, epoch 等），不同 lambda 值導致的精度結果，此模型共選用 37 個特徵值。

Lambda 值	Accuracy (training set)	Accuracy (validating set)
0	0.8449	0.8548
0.01	0.8449	0.8483
0.0001	0.8458	0.8434

由表格可以觀察出，面對不同的 lambda 值，所得到的精度差異並不大，由此可推得模型採用的 37 個特徵，並無 overfitting 的情況發生，故有無加上限制都不影響最後模型的表現。

5.

以相關係數(取絕對值)來看，前五名的參數分別為：

('marital_status_ Married-civ-spouse', 0.4446961), ('relationship_ Husband', 0.40103526), ('education_num', 0.3351539), ('marital_status_ Never-married', 0.318440), ('age', 0.234037)

最高值僅 0.4446，可見各特徵皆無顯著的線性相關，且當面對不同的參數組合時，其 correlation 也會有所不同；我認為非數值資料中並無特別顯著的因子，若將各資料與 label 關係以直方圖表示，可發現於該資料表現為真(value=1)的人數在兩種分類可能是差不多的，推測原因可能是由於(>50k)的資料過少，縱使比例較高，但演算法仍無法增加其權重。歸咎於資料集分類的比例不均（其中一類過多），我認為數值資料的影響相對來說較大，其中 ***capital_loss*** 的影響較大，約可提升 0.1~0.2 的訓練精度，雖然其大部分的值為 0，但撇除這些，在有值的部分，兩類的分佈是有所差異，這樣的特性能有效地讓演算法朝著較好的訓練精度去修正，最終在 Kaggle 上的結果也較佳。