學號：b05901070 系級：電機四 姓名：蔡昌廷

1. (2%) 請比較實作的 generative model 及 logistic regression 的準確率，何者較佳？請解釋為何有這種情況？

   在驗證 data 的部分，為了求公平起見，training data 是直接用整個 X_train 去 train，另外 testing 結果皆為上傳 Kaggle 後的準確率，結果如下表所示。

| | generative | logistic |
|---|---|---|
| traing | 0.87218 | 0.88611 |
| testing | 0.88212 | 0.88907 |

   由上表可看出本次作業在 logistic regression 的準確率較高，會有此現象的原因在於 generative model 在一開始會假設每一個 data 是來自一個機率模型，然而在 data 數量夠多時，或許根本不存在這個假設，相較之下 logistic 是根據現有 data 去擬合出最佳的 function，在此次實作上較為合適。

2. (2%) 請實作 logistic regression 的正規化 (regularization)，並討論其對於你的模型準確率的影響。接著嘗試對正規項使用不同的權重 (lambda)，並討論其影響。(有官 regularization 請參考 https://goo.gl/SSWGhf p.35)

   以下結果為針對 logistic regression 做 L2 正規化的結果，λ值由 0.01 至 10

| | Training loss | Training Accuracy |
|---|---|---|
| λ = 0.01 | 0.265579 | 0.886077 |
| λ = 0.1 | 0.265577 | 0.886113 |
| λ = 0 | 0.265579 | 0.886113 |
| λ = 1 | 0.265723 | 0.885966 |
| λ = 10 | 0.270791 | 0.883754 |

   從上表結果發現實作正規化並無使 model 表現更好，推測原因可能是 model 並沒有 overfitting 情形，因此做正規化效果不大。

3. (1%) 請說明你實作的 best model，其訓練方式和準確率為何？

   此次 best model，因為從原始 data 去分析，認為最後一項資料(years)對結果感覺並無顯著影響，因此我先將 X_train 最後兩個資料拿掉，接著我針對 X_train 裡面非 one-hot encoding 的資料做三次多項式的回歸($b + w1*x + w2*x^2 + w3*x^3$)，最後讓 model 跑 20000 次 iteration 得出來的結果，另外有使用 Adam 作為 optimizer，最後上傳 Kaggle 的準確率為 0.89283。

4. (1%) 請實作輸入特徵標準化 (feature normalization)，並比較是否應用此技巧，會對於你的模型有何影響。

| | 有 normalization | 無 normalization |
|---|---|---|
| Training loss | 0.264394 | 2.826281 |
| Training accuracy | 0.886401 | 0.813413 |
| Validation loss | 0.283725 | 2.938249 |
| Validation accuracy | 0.877441 | 0.808883 |

   此題我先將整個 training set 以 9:1 的形式分成 training set 跟 validation set，由上表結果可發現在有做 normalization 的情況下，整個 model 的表現好非常多，最主要原因在於 normalization 可以將不同的資料的 scale 以同一個標準做轉換，在做 gradient descent 時可以優化其下降的結果，還可以提高準確度。