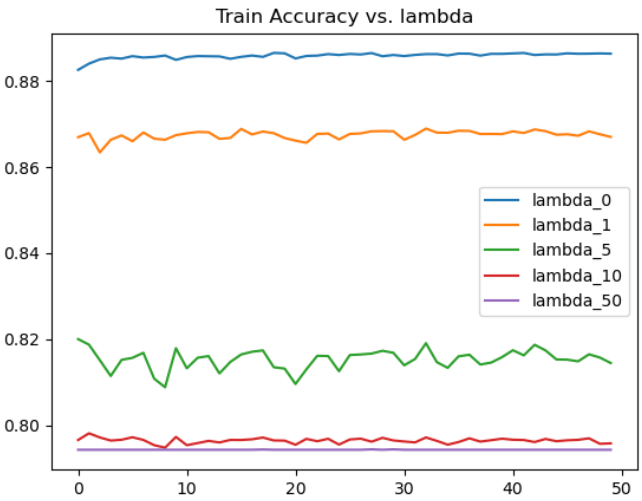
Machine Learning HW2

學號:B06902060　系級:資工三　姓名:鄒宗霖

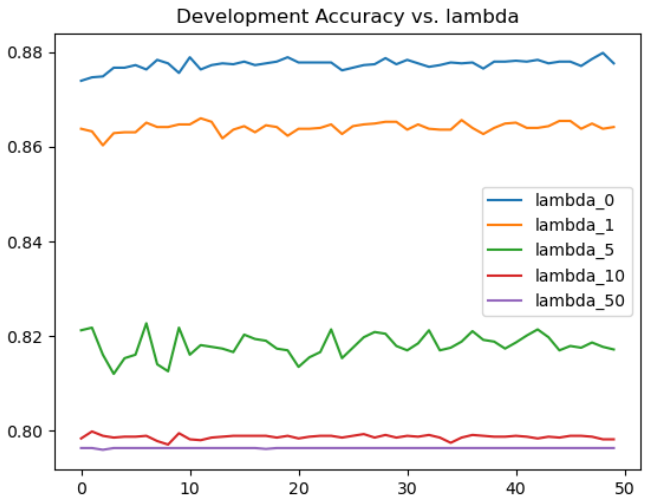
1. (2%) 請比較實作的 generative model 及 logistic regression 的準確率，何者較佳? 請解釋為何有這種情況?

   根據實驗的結果，logistic regression 的準確率比 generative model 來的高，在 Kaggle 上得到的分數分別是 0.88617 以及 0.87857，可能的原因是 generative model 假設了 P(x | C) (probability from class) 來自於某個 Gaussian Distribution，這樣的好處是擁有較少的 data 也可以有不錯的表現，並且不容易受到錯誤資料的影響，但 P(x | C) 可能不是來自於某個 Gaussian Distribution，當資料量足夠且資料正確性高的情況下，generative model 失去了這些優勢，表現的比 logistic regression 還差。

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。(有關 regularization 請參考 https://goo.gl/SSWGhf p.35)

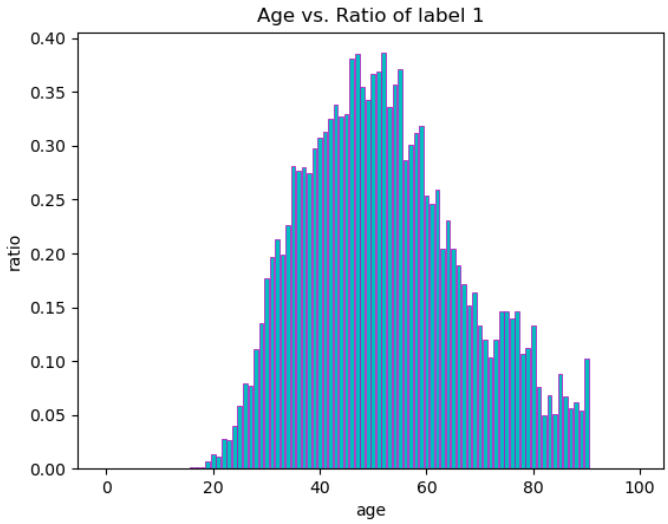


   以下為實作 logistic regression 正規化的結果，我使用了五種不同的 lambda 分別是 0, 1, 5, 10, 50，依照第一張圖來看，lambda 越高會導致 training data 的準確率降低，可能的原因是當 lambda 逐漸上升時，增加了考慮 model 是否平滑 (weight 總和越小越平滑) 的比重，而減少了考慮 training error 的比重。



然而依照第二張圖來看，lambda 越高也會導致 development data 的準確率降低，可能的原因是當 lambda 逐漸上升時，漸漸地把一些 weight 總和太大的 model 排除了，但排除掉的 model 可能是比較好的 model。

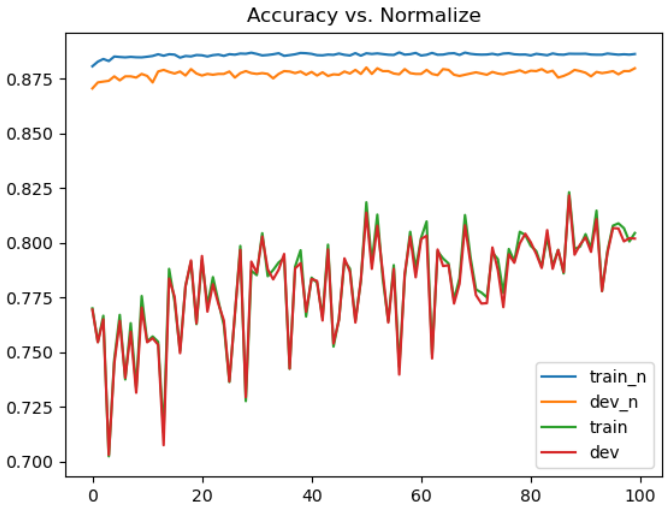
3. (1%) 請說明你實作的 best model，其訓練方式和準確率為何?



	step 1	step 2	step 3	step 4
Accuracy	0.89016	0.89030	0.89059	0.89305

一開始的 model 準確率為 0.88617，step 1 就是增加 iteration (iter = 50)，因為 iteration 如果太小可能參數都還沒有收斂就停止更新參數了，準確率為 0.89016；step 2 就是把一些比較不重要的參數拿掉，留下 450 個 features，準確率為 0.89030；step 3 就是把年齡這個欄位的值改成該年齡年收入超過五萬美元的比例，如上圖所示，準確率為 0.89059；step 4 就是手動新增了幾個 features，(如 : 是否有 capital gains、是否有 capital losses、是否有 dividends from stocks、是否 weeks worked in year 大於零)，準確率為 0.89305。

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。



上圖為實作特徵標準化的結果，我們可以看到經過特徵標準化的 model 在第一個 iteration 之後就有很高的準確率，並且穩定及緩慢的提升；沒有經過特徵標準化的 model 雖然在歷經多個 iteration 之後準確率提升了不少，但在過程中準確率的起伏變化很大，比起經過特徵標準化的 model 準確率還是低了許多。可能的原因是經過 feature normalization 之後，每一次 gradient descent 更新參數的方向都是指向 loss function 的最低點 (如 : 兩個參數在二維平面上 loss function 的等高線是同心圓)，不但可以提高參數的收斂速度，還可以使得每一個 feature 對 loss function 的影響程度相同 (平均值都是 0、標準差都是 1)；然而沒有經過特徵標準化的 model，每一個 feature 的分散程度差異甚大，loss function 的等高線圖可能扭曲變形，導致 gradient descent 更新參數的方向很可能不是指向 loss function 的最低點，因此更新參數過程中準確率的起伏變化很大，另外，每一個 feature 對 loss function 的影響程度不同，可能會造成 model 的結果失真 (如 : w1 -> [0, 1] , w2 -> [0, 10000]，使得 w2 的影響力可能大於 w1，但實際上 w1 的指標意義大於 w2)。