Homework 2 Report - Income Prediction

學號：r05323040　系級：經濟所碩二　姓名：田家駿

1. (1%) 請比較你實作的 generative model、logistic regression 的準確率，何者較佳？

Ans.上傳到 Kaggle 上的 Public 成績 generative model 為 0.83，而 logistic regression 為 0.84606，logistic regression 因無假設資料的分配，在預測的表現上較 generative model 表現較好。

2. (1%) 請說明你實作的 best model，其訓練方式和準確率為何？

Ans.我的 best model 是使用 svm，因 svm 去解三萬多筆資料效率很差，因此我造了很多 svm 模型，每次都只 fit 6000 筆資料，最後再用多數決的方式決定我的預測，上傳準確率為 0.85909，比 logit 好了 0.1 左右，另外也試了用高斯 kernel 然而表現並不理想，最後只使用 linear kerne。

3. (1%) 請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。

Ans. 這裡我使用的是 minmax 標準化，把連續型的變數壓到 0 到 1 區間，沒有做 normalize kaggle 成績為 0.8081，learning rate 為 0.0000001，作完 normalize，分數就上升到 0.84606，learning rate 為 0.01，可以看到做完 normalize 之後預測表現便比較好。

4. (1%) 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。

Ans.我使用了 L2 norm 的 regularization，CV tune 出來的的參數為 0.5，上傳 kaggle 分數為 0.8469，稍微比原本模型表現好一點，另外我又上傳了參數為 10 以及 100，10 的成績為 0.8390，100 的成績為 76.05，可以看到雖然參數越大，模型會越平緩，但預測表現也可能會變差。

5. (1%) 請討論你認為哪個 attribute 對結果影響最大？

Ans.我認為年紀的影響會是最大的，可以看到越往中年 outcome 實現的比例較高，但隨著年紀的增加，outcome 比例卻降低，因此我考慮了年紀二次項，把兩項年紀變數拉掉作一次 logistic，上傳分數為 0.8424，可以看到，考慮了年紀，能讓我們 outsample 的表現較好一點。