學號：R06725005 系級： 資管碩一 姓名：郝思喬

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

答：

|  |  |  |
| --- | --- | --- |
|  | Private Score | Public Score |
| Generative Model | 0.84227 | 0.84520 |
| Logistic Regression | 0.84964 | 0.85552 |

Logistic Regression的準確率較佳。

2.請說明你實作的best model，其訓練方式和準確率為何？

答：

Best Model為使用gradient boosting的方式來實作，訓練方式為使用Scikit-learn 0.19.0套件中的gradient boosting來實作，並且設定depth=5。

|  |  |  |
| --- | --- | --- |
|  | Private Score | Public Score |
| Gradient Boosting | 0.87372 | 0.87874 |

3.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。

答：

|  |  |  |
| --- | --- | --- |
|  | 標準化前 | 標準化後 |
| Generative Model | 0.84227 | 0.84227 |
| Logistic Regression | Overflow | 0.84964 |
| Gradient Boosting | 0.87372 | 0.87372 |

Generative Model - 有沒有標準差都沒差別。

Logistic Model - 沒作標準化會overflow無法實作。

Gradient Boosting - 有沒有標準差都沒差別。

4. 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。

答：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | λ = 100 | λ = 10 | λ = 1 | λ = 0.1 | λ = 0.01 |
| Accuracy | 0.84166 | 0.84166 | 0.84160 | 0.85057 | 0.80650 |

在λ = 0.1時會較好

5.請討論你認為哪個attribute對結果影響最大？

根據sigmoid與likeness function，w絕對值越大的attribute影響越大。

capital\_gain越高，年收越會高於50K。

工作經驗是Never-worked的，年收越會低於50K。

Without-pay越高，年收越會低於50K。

學歷是Preschool的年收越會低於50K。

住在Holand-Netherlands的年收越會低於50K。