銀行客戶的流失預測

統計111 蔡儀萱 H24076150

統計111 林少穎 H24071273

．競賽敘述與目標

我python的機器學習與應用，並且利用自行上網搜尋或老師上課教的東西，學我習使用scikit-learn的套件，來分析銀行客戶的流失預測。

．資料前處理

我沒有採用此三個特徵RowNumber、Customerld、Surname

．特徵處理與分析

我1. 將Geography前面的S去除

我2. 將Gender中，Male換成0，Female換成1

我3. 將Age中的年齡分群(年齡//4 or 年齡//5)

我4. 將Credit Score、Balance、Estimated Salary做標準化或一般化

．預測訓練模型

我1.KNN

參數：n\_neighbors=9，p=1，weights='distance'

※在進行KNN訓練與預測前，先將資料做transform

我2.RandomForestClassifier

參數：n\_estimators=200，oob\_score=True，n\_jobs=-1，random\_state=1，max\_features='auto'，min\_samples\_leaf=10

我3.MLPClassifier

參數：hidden\_layer\_sizes=(50,50)，max\_iter=200，alpha=1e-4，solver='sgd'，verbose=10，tol=1e-4，random\_state=1，learning\_rate\_init=0.1，momentum=0.5，learning\_rate='adaptive'

我4.GuassianNB

參數：無更改(使用預設值)

．預測結果分析

▼將Credit Score，Balance，Estimated Salary做標準化、一般化的結果

|  |  |  |  |
| --- | --- | --- | --- |
| KNN | Standardize | Normalize | Standardize & Normalize |
| Accuracy | 0.8400 | 0.8430 | 0.8410 |
| Precision | 0.6977 | 0.7135 | 0.7042 |
| F Score | 0.4839 | 0.4919 | 0.4854 |
| Final | 0.6549 | 0.6637 | 0.6577 |

|  |  |  |  |
| --- | --- | --- | --- |
| MLPClassifier | Standardize | Normalize | Standardize & Normalize |
| Accuracy | 0.8580 | 0.8580 | 0.8575 |
| Precision | 0.7763 | 0.7763 | 0.7804 |
| F Score | 0.5449 | 0.5449 | 0.5396 |
| Final | 0.7082 | 0.7082 | 0.7072 |

|  |  |  |  |
| --- | --- | --- | --- |
| GuassianNB | Standardize | Normalize | Standardize & Normalize |
| Accuracy | 0.8310 | 0.8310 | 0.8295 |
| Precision | 0.7445 | 0.7445 | 0.6633 |
| F Score | 0.3764 | 0.3764 | 0.4326 |
| Final | 0.6232 | 0.6232 | 0.6209 |

▼調整各模型預測結果所佔的比例

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| KNN:Forest:MLP:Bayes | 1:1:1:1 | 1:2:2:1 | 1:5:4:1 | 1:7:5:1 | 2:12:9:1 |
| Accuracy | 0.8555 | 0.8575 | 0.8630 | 0.8625 | 0.8625 |
| Precision | 0.8537 | 0.8181 | 0.8047 | 0.8037 | 0.8037 |
| F Score | 0.4921 | 0.5194 | 0.5581 | 0.5557 | 0.5557 |
| Final | 0.7026 | 0.7108 | 0.7235 | 0.7222 | 0.7222 |

※可發現，RandomForestClassifier需佔最多比例，但若比例達0.5，並不會提高整體分數，而我們試到最好的結果為1:5:4:1，其分數為0.7235。