國立雲林科技大學

資訊管理研究所

111學年度第二學期

機器學習

專案作業一

指導教授：許中川教授

組　　員：M11123052 賴俊佑

M11123062 陳靖穎

M11223002 陳怡君

摘要

現今物價飛漲的時代，對於成本考量下如何協助尋找有利的客戶成為討論的議題。。

本研究想要預測客戶是否訂閱了定期存款，利用Keras 建立一個前饋神經網路模型，並將資料分成類別類型欄位以及數值型欄位，類別類型欄位Precision, Recall及F1指標，數值預測使用MAE、MAPE與RMSE指標。各類別分別採三組不同的差異做訓練，類別型在第三組訓練最好，數值型則是在第二組效果比較好。。

關鍵字：前饋神經網路、顧客分析、機器學習

一、緒論

現在社會來到了萬物齊漲的時代，看到許多店家公告上漲的消息。根據行政院主計總處(主計總處，2023)的資料顯示消費者物價指數年增率來到了2.43%，而基本工資也在2023年1月1日起調漲，月薪從現行2萬5,250元調增至2萬6,400元，調幅為4.56%，時薪的部分也從168元調增至176元，調增8元(行政院，2022)，因此對於公司來說，資源分配上需要更有效的利用。

1.1 動機

如今公司成本不斷上升的情況下，成本的計算也變得很重要。在電話訪談需要花費人力去對客戶做電話訪問，與其用灑網式的電訪提問，還不如利用現有的客戶資料來找到潛在有價值的客戶。

1.2 目的

為了找到適合的訓練資料，分別對資料的類別類型欄位以及數值型欄位，類別類型欄位使用Precision, Recall及F1指標，數值預測使用MAE、MAPE與RMSE指標，找出最適合的模型，以利了解客戶是否有訂閱定期存款。

二、方法

本研究預計分類別型跟數值型欄位兩種預測類別。匯入資料後先進行資料預處理，將day跟month數值去除，去除後再將資料分組並模型資料，接著進行不同的時期(Epochs)跟批次(Batch Size)分析。類別型資料採用Precision, Recall及F1績效衡量指標；而數值型則採用Mean Absolute Error (MAE)、Mean Absolute Percentage Error (MAPE)、Mean Squared Error (MSE)。

三、實驗

3.1 資料集

* 資料集名稱：Bank Marketing
* 資料筆數：45211
* 屬性數量：17

**表1** *Bank Marketing資料集屬性簡介*

|  |  |  |
| --- | --- | --- |
| 屬性名稱 | 型態 | 尺度 |
| age | Int | 比例 |
| job | String | 區間 |
| marital | String | 區間 |
| education | String | 順序 |
| balance | Int | 比例 |
| housing | String | 名目 |
| loan | String | 名目 |
| contact | String | 名目 |
| day | Int | 區間 |
| month | Int | 區間 |
| duration | Int | 比例 |
| campaign | Int | 比例 |
| pdays | Int | 比例 |
| previous | Int | 比例 |
| poutcome | String | 區間 |
| y | String | 名目 |

3.2 前置處理

本研究類別型預測前處理步驟為將day及month資料刪除，再提取類別型資料，將類別值轉換為整數編碼再拆成特徵集合（X）跟測試集（y），測試集大小設定為總資料的15%，接著對資料進行標準化以利提高模型的性能及穩定性。

數值型預處理跟類別型類似，差別在將資料集中的 age 欄位從特徵中刪除，並將刪除後的特徵存入變數 X 中，將age欄位的值存入目標變數 y 中。接著，兩個類型都使用 train\_test\_split 函數將資料集分成訓練集和測試集，再對資料進行正規化。

3.3 實驗設計

使用 Keras 建立一個前饋神經網路模型，進行訓練和評估。

類別型預測模型有一個有 14 個輸入節點和 1 個輸出節點的隱藏層、一個具有 20 個神經元的隱藏層(relu)，以及一個輸出層，設定損失函數為二元交叉熵（binary\_crossentropy）、優化器使用Adam，並設定模型的評估指標為精確度（precision\_m）、召回率（recall\_m）和 F1 值（f1\_m）。首先先採用訓練 5 個時期(epochs = 5)，每個批次包含 100 個樣本(batch\_size = 100)，第二組採訓練 10 個時期(epochs = 10)，每個批次包含 100 個樣本(batch\_size = 100)，第三組採訓練 5 個時期(epochs = 5)，每個批次包含 50 個樣本(batch\_size = 50)。

數值型預測模型則設定損失函數為MSE，並設定評估指標為 MAE、 MAPE、 MSE。

3.4 實驗結果

**表2** *類別型各架構輸出值*

|  |  |
| --- | --- |
| 架構 | 輸出值 |
| epochs=5, batch\_size=100 |  |
| Precision | .201 |
| Recall | .073 |
| F1 | .101 |
| epochs=10, batch\_size=100 |  |
| Precision | .260 |
| Recall | .108 |
| F1 | .141 |
| epochs=5, batch\_size=50 |  |
| Precision | .379 |
| Recall | .159 |
| F1 | .210 |

**表3** *數值型各架構輸出值*

|  |  |
| --- | --- |
| 架構 | 輸出值 |
| epochs=5, batch\_size=100 |  |
| MAE | 9.951 |
| MAPE | 24.507 |
| MSE | 184.530 |
| epochs=10, batch\_size=100 |  |
| MAE | 7.970 |
| MAPE | 26.592 |
| MSE | 98.137 |
| epochs=5, batch\_size=50 |  |
| MAE | 8.287 |
| MAPE | 21.363 |
| MSE | 104.409 |

四、結論

透過演算法結果可以發現，以類別型的預測來說第三組採訓練 5 個時期(epochs = 5)，每個批次包含 50 個樣本(batch\_size = 50)的輸出結果表現最好。數值型的預測中，第二組採訓練 10 個時期(epochs = 10)，每個批次包含 100 個樣本(batch\_size = 100)這組的總體結果比較好，MAE跟MSE都有不錯的表現，但MAPE表現在第二組最差，第三組表現最好。。

參考文獻

行政院主計總處（2023）。*消費者物價指數年增率*。2021年 01月27日，取自<https://www.stat.gov.tw//Point.aspx?sid=t.2&n=3581&sms=11480>

行政院（2022）。*核定基本工資調漲方案 蘇揆：盡全力照顧勞工，並讓勞資共榮共好*。行政院新聞傳播處，取自<https://www.ey.gov.tw/Page/9277F759E41CCD91/97672e16-d759-4d58-8f9b-f9805305be7e>

摘要

在大數據時代，政府如何快速的統整資料，甚至是進一步預測資料，超前部屬已成為每個政府單位必備的能力。

本研究利用Keras前饋神經網路訓練，欲預估出每週的工作時數，透過不同的時期跟權重調整下，本研究發現在第二組訓練中評估指標Mean Absolute Error (MAE)、Mean Absolute Percentage Error (MAPE)、Mean Squared Error (MSE)表現最好。

關鍵字：前饋神經網路、資訊分析、機器學習

一、緒論

現在隨著大數據時代的興起以及政府電子化作業逐漸普及，各項數據已經有專門的平台來顯示。而全台平均薪資已經成為大家每年都會比較的項目之一，而政府如何快速的產生資料，或是提前預測明年趨勢成為政府新的課題。

1.1 動機

自從中央疫情指揮中心記者會中出現了「超前部屬」這個詞後(求真百科，2020)，各政府單位都掀起一股「超前部屬」的風潮來顯示單位對於未來業務方向已有萬全的準備。而為了保障勞工的權益，每週的工作時數需要嚴格的把關，提早預測工作時數的變化可以使政府單位提早因應相關的政策調整。

1.2 目的

為了準確的預測每週的工作時數，本研究利用了Keras前饋神經網路進行資料集的訓練，希望了解在哪樣的訓練數期及權重調整下有較好的預測結果。

二、方法

本研究預計使用adult.data當作訓練資料集，adult.test當測試資料集，在執行完資料前處理及正規化後便開始訓練。訓練分成三組不同的方法訊練，分別對於訓練時期及權重計算不同的調整。

三、實驗

3.1 資料集

* 資料集名稱：Adult
* 資料筆數：48842
* 屬性數量：14

**表1** *adult資料集屬性簡介*

|  |  |  |
| --- | --- | --- |
| 屬性名稱 | 型態 | 尺度 |
| age | Int | 比例 |
| workclass | String | 名目 |
| fnlwgt | Int | 比例 |
| education | String | 順序 |
| educational-num | Int | 區間 |
| marital-status | String | 名目 |
| occupation | String | 名目 |
| relationship | String | 名目 |
| race | String | 名目 |
| gender | String | 名目 |
| capital-gain | Int | 比例 |
| capital-loss | Int | 比例 |
| hours-per-week | Int | 比例 |
| native-country | String | 名目 |
| income | String | 順序 |

3.2 前置處理

把資料匯入後，先刪除欄位內的缺失值，避免影響模型訓練。接著鎖定研究所需要的欄位內容是名目的話將其轉為數值，並針對訓練要的「hours-per-week」欄位進行測試集跟資料集切割，最後進行正規化。

3.3 實驗設計

使用 Keras 建立一個神經網路模型進行訓練和評估。有一個 14 個輸入節點和 1 個輸出節點的隱藏層、一個具有 20 個神經元的隱藏層，以及一個輸出層，設定損失函數為Mean Squared Error (MSE)、優化器為 Adam，並設定模型的評估指標Mean Absolute Error (MAE)、Mean Absolute Percentage Error (MAPE)、Mean Squared Error (MSE)。首先先採用訓練 5 個時期(epochs = 5)，每個批次包含 100 個樣本(batch\_size = 100)，第二組採訓練 10 個時期(epochs = 10)，每個批次包含 100 個樣本(batch\_size = 100)，第三組採訓練 5 個時期(epochs = 5)，每個批次包含 50 個樣本(batch\_size = 50)。

3.4 實驗結果

**表2** *預測結果*

|  |  |
| --- | --- |
| 架構 | 輸出值 |
| epochs=5, batch\_size=100 |  |
| MAE | 10.907 |
| MAPE | 35.509 |
| MSE | 198.359 |
| epochs=10, batch\_size=100 |  |
| MAE | 7.688 |
| MAPE | 29.995 |
| MSE | 131.812 |
| epochs=5, batch\_size=50 |  |
| MAE | 7.929 |
| MAPE | 30.884 |
| MSE | 132.456 |

四、結論

透過預測結果可以發現，第二組採訓練 10 個時期(epochs = 10)，每個批次包含 100 個樣本(batch\_size = 100)這組的總體結果相較於其他兩組有比較好的表現，對於此資料集的話可能選用第二組結果較為適合。

參考文獻

求真百科（2020）。*超前部屬*。取自<https://factpedia.org/index.php?title=%E8%B6%85%E5%89%8D%E9%83%A8%E7%BD%B2&variant=zh-hant>