元　　智　　大　　學

資 訊 工 程 學 系 (所)

大數據資料分析　論　文

員工離職預測

研 究 生：韓志鴻、邱奕鋒、張羿翔

指導教授：簡廷因

中 華 民 國　一一三　年　六　月

目　錄

表　　目　　錄

圖　　目　　錄

第一章 摘要

本研究旨在使用多種機器學習模型對員工離職行為進行預測。使用來自 AIdea 人工智慧共創平台的數據集，我們進行了資料清洗、特徵篩選及模型訓練，並比較了不同模型的預測性能。我們採用了深度神經網絡（DNN）、XGBoost、隨機森林（Random Forest）、決策樹（Decision Tree）和單類支持向量機（OneClass SVM）進行實驗。實驗結果顯示，DNN 模型在 F1 得分上表現最佳，為 0.2723988，在 602 個參賽者中排名第 3。本文進一步通過 SHAP 分析各模型的重要特徵，並探討模型性能差異的原因，為未來的模型改進提供了重要參考。

第二章 介紹

員工離職預測是企業人力資源管理中的重要課題，準確預測員工的離職意圖有助於企業制定有效的員工保留策略。本研究旨在使用多種機器學習模型對員工離職行為進行預測，並比較各模型的預測性能。數據來源於 AIdea 人工智慧共創平台，包含 14392 筆訓練資料和 3739 筆測試資料，每筆資料包含 47 個特徵，如性別、職等、工作資歷、升遷速度等。我們使用了 DNN、XGBoost、Random Forest、Decision Tree 和 OneClass SVM 等五種模型進行預測，並通過 SHAP 分析各特徵對模型預測結果的影響，探討模型性能差異的原因。

第三章 研究方法

一、資料描述

1. 資料來源：AIdea 人工智慧共創平台
2. 資料集內容：
   - train.csv 有 14392 筆資料、test.csv 有 3739 筆資料，皆 47 個欄位。欄位特徵包含如性別、職等、工作資歷、升遷速度...等特徵。
   - 其中的 PerStatus 欄位為預測目標，0 代表未離職；1 代表離職。

二、資料分析與處理

1. 透過 Rshiny 來呈現一個網頁式介面。

2. 「特徵之間的離職佔比」和「特徵值資料佔比」頁面可呈現個特徵資料的佔比分布。

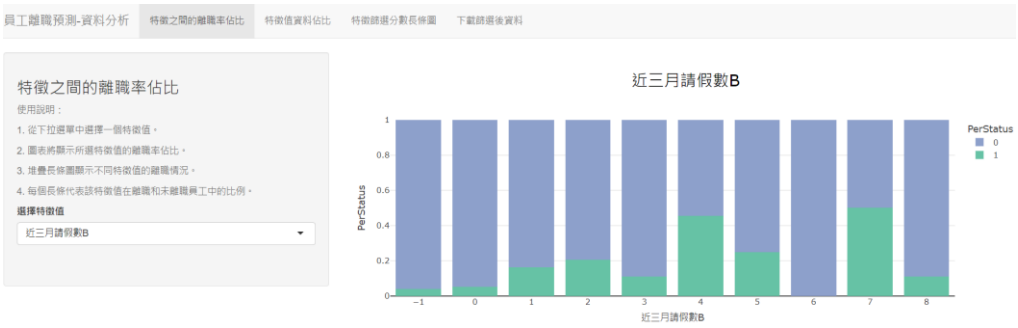

圖 1、特徵之間的離職佔比

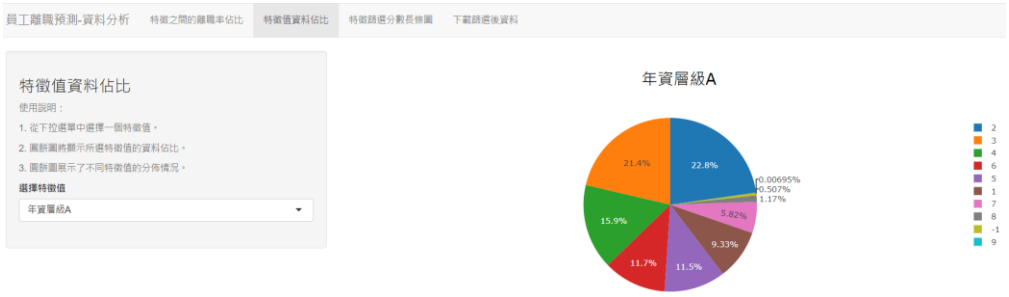

圖 2、特徵值資料佔比

3. 「特徵篩選分數長條圖」頁面通過變異數分析（ANOVA）計算各特徵的 p 值，篩選出對預測目標 PerStatus 可能具有重要影響的特徵。p 值越小，表示該特徵與預測目標的相關性越高，因此更適合用於後續的模型訓練。此頁面提供了一個可選擇的 p 值範圍（0 到 0.1），並顯示所有在此範圍內的特徵。這些特徵按 p 值從小到大的順序排列，並以長條圖的形式展示，方便觀察哪些特徵適合用於模型訓練（此特徵篩選在 Python 上的實作相當於使用 sklearn 中的 SelectKBest 函數）。

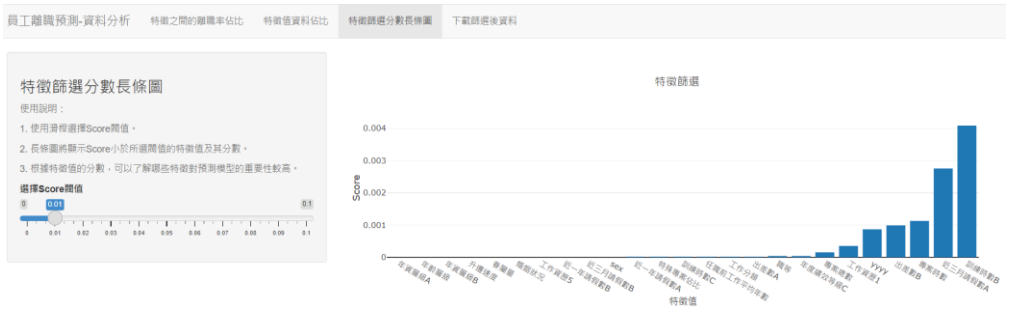

圖 3、特徵篩選分數長條圖

4. 「下載篩選後的資料」頁面，可透過滑桿選擇 p 值範圍，並將篩選後的特徵顯示在下方的多選欄位，後續亦可在此多選欄位增減其他特徵，並將篩選後的資料集顯示在右側，最後可下載該資料集作為後續的模型訓練。

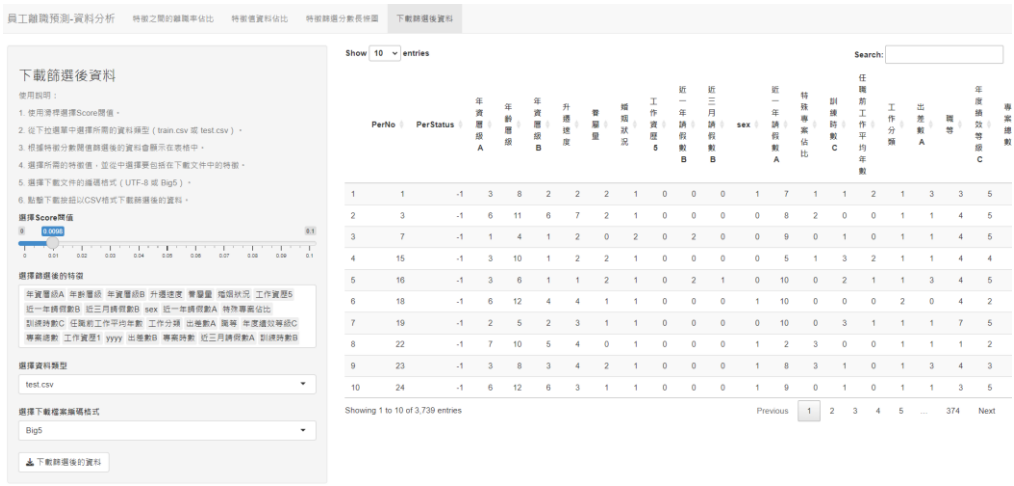

圖 4、下載篩選後的資料

三、模型訓練與預測

1. 從 Rshiny 上下載經處理的資料集 processed_train.csv 和 processed_test.csv。

2. 由於會有資料不平衡的問題，因此對訓練集使用 SMOTE 來合成少數樣本。

3. 將處理後的資料投入模型訓練，並比較各模型的訓練結果。

4. 模型訓練完成後，預測 processed_test.csv 內每位員工(PerNo)是否會離職，並將其結果上傳至 AIdea 官網來評估分數與排名。

5. 最後會對每個模型使用 SHAP 分析每個因子的貢獻，來衡量每個特徵對模型預測的貢獻程度。

第四章 實驗結果

一、實驗結果分析

我們選擇了五種機器學習模型進行實驗，分別是深度神經網絡（DNN）、XGBoost、隨機森林（Random Forest）、決策樹（Decision Tree）和單類支持向量機（OneClass SVM）。各模型的 F1 得分和在 AIdea 中的排名如下表所示：

Model	F1 Score	Ranking
DNN	0.2731884	3/602
XGBoost	0.1413043	375/602
Random Forest	0.1358328	391/602
DecisionTree	0.1342121	396/602
oneclass SVM	0.0760233	508/602

表 1、實驗結果

其中 DNN 模型以 F1 得分 0.2723988 在所有參賽者中排名第 3。對於員工離職預測有較好的效能，且為本次實驗下來最好的結果。相比之下，XGBoost、Random Forest、DecisionTree 以及 OneClass SVM 的表現則相對較弱，分別的 F1 Score 為 0.141、0.136、0.134、0.076。接下來我們根據 SHAP 分析這些模型的重要性分佈，並探討結果較差的原因。

二、分析模型

為了進一步理解各特徵對模型預測結果的影響，我們使用 SHAP 對各模型進行了特徵貢獻度分析。以下是各模型的 SHAP 分析結果：

1. DNN（Deep Neural Networks）

根據圖一 SHAP 分析顯示，年資層級 B、性別、年度績效等級 C、年資層級 A 等特徵對預測結果影響較大。其中像是「年資層級 B」、「年資層級 A」的特徵在 DNN 中有顯著的影響，可能是因為擁有中等年資的員工在離職行為上較大變化，抑或是年資較短的員工可能還未對公司產生足夠的歸屬感，導致離職率較高。

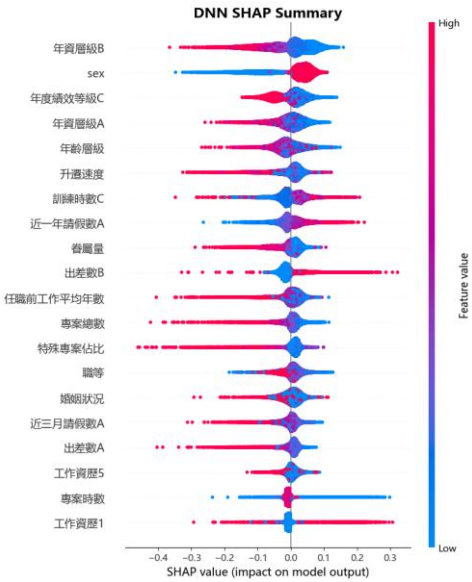

圖 5、SHAP 分析-DNN

而我們認為 DNN 在本次實驗中有良好的表現在於其強大的非線性建模能力和特徵自動提取能力，使它能夠從複雜的員工離職資料中學習有效的特徵模式。但由於其模型結構複雜，訓練時間較長，且需要大量數據以避免過擬合。

2. XGBoost

根據圖二SHAP分析顯示，年資層級B、訓練時數C、特殊專案佔比、出差數A、年資層級A等特徵對預測結果影響較大。與DNN不同的是在「訓練時數C」可能表現出訓練投入與員工離職之間有潛在關聯，或是「特殊專案佔比」可能反映員工的工作負荷和滿意度。

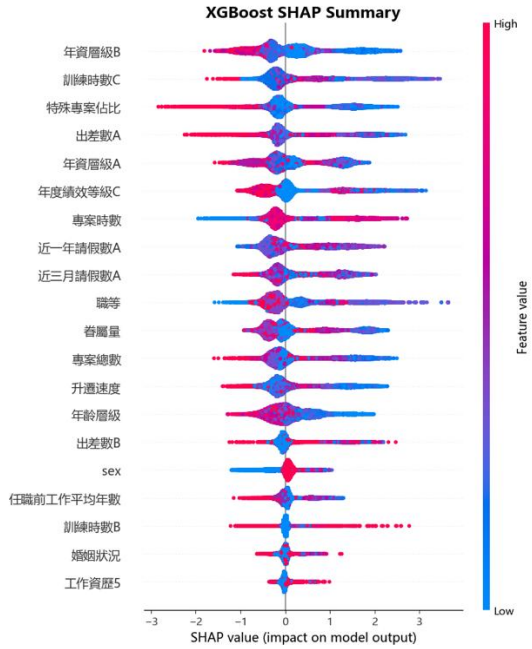

圖 6、SHAP 分析-XGBoost

而我們認為 XGBoost 沒有較好的結果，可能是這些特徵選擇不夠恰當，導致模型無法有效學習和預測。抑或是在訓練時參數調整不當，可能導致模型過擬合或欠擬合，預測性能會受到影響。也許後續可以試著用其他特徵或是訓練參數來調整。

3. Random Forest

根據圖二 SHAP 分析顯示，年資層級 B、年資層級 A、升遷速度、近一年請假數 A、年齡層級等特徵對預測結果影響較大。與 DNN 不同的是在「升遷速度」可能在升遷速度反映員工在公司內的職業發展情況，對離職決策有重要影響。或是在「近一年請假數 A」在請假頻率可能反映員工的工作壓力和滿意度，進而影響其離職意願。

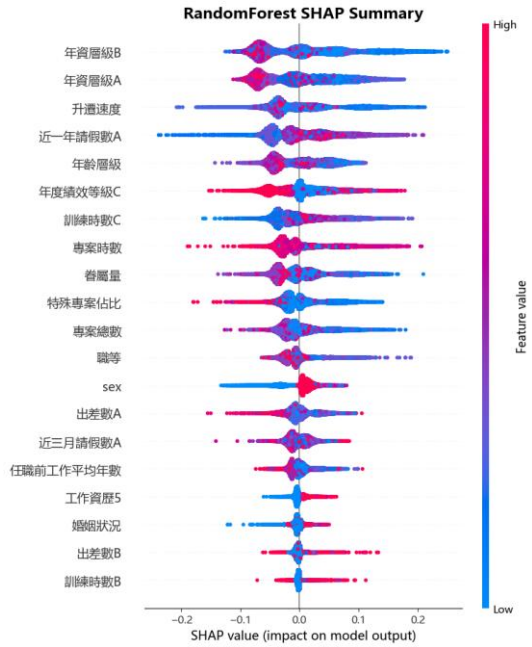

圖 7、SHAP 分析- Random Forest

而我們認為 Random Forest 對於高維特徵和非線性關係具有較強的處理能力，但對於離職預測這類數據分佈可能較偏的問題，表現可能不如 DNN 這類能夠學習更複雜模式的模型。也可能在模型參數的選擇和調整需要加強與改善，才能提升整體的準確度。

4. DecisionTree

根據圖二 SHAP 分析顯示，年資層級 B、專案時數、訓練時數 C、近一年請假數 A、升遷速度等特徵對預測結果影響較大。與 DNN 不同的是在「專案時數」可能在參與專案的時數可能反映員工的工作負荷和滿意度。

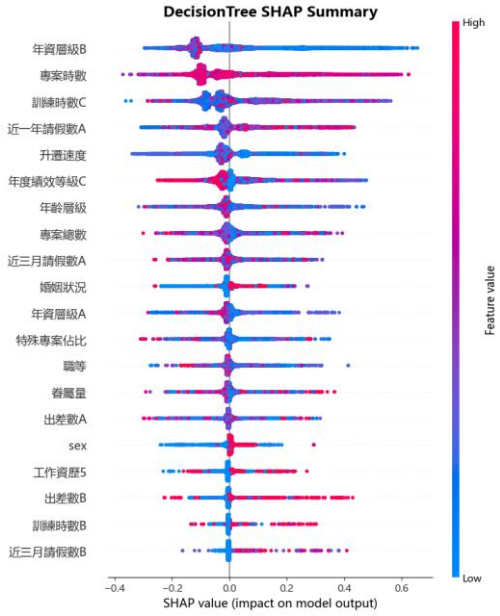

圖 8、SHAP 分析- DecisionTree

我們認為雖然 Decision Tree 直觀且易於解釋，但其對於數據波動和噪音較敏感，容易過擬合。且 Decision Tree 模型穩定性和泛化能力較差，導致其表現不如其他複雜模型。

5. oneclass SVM

我們認為該模型適合異常檢測，但在多類分類問題上可能表現不佳。假設數據主要集中在一個類別，而對於離職預測這類二元分類問題，該模型的適用性較差，導致 F1 Score 和排名較低。

第五章 結論

在本次實驗中，DNN 模型因其強大的非線性建模能力和特徵自動提取能力，在員工離職預測任務中表現最佳。XGBoost 和 Random Forest 也在一定程度上能夠捕捉數據中的變異，但在處理高維數據和避免過擬合方面相對 DNN 有所欠缺。Decision Tree 和 OneClass SVM 在簡單性和異常檢測方面各有優勢，但在多變因的員工離職預測中表現相對較弱。SHAP 分析結果顯示，特徵選擇對模型性能有重要影響，為未來的模型改進提供了關鍵方向。