

Aldea

# 員工離職預測

1093304 韓志鴻、1093324 邱奕鋒、1093345 張羿翔

# Introduction



# Introduction

## 研究主題

- **Employee Resignation Prediction**

- **人才管理的重要性**：人才是企業最重要的資源，預測員工離職傾向的目的是為了留住優秀人才，確保企業持續成長。
- **離職預測的方法**：使用大數據與人工智慧技術來分析員工的離職風險，進而針對風險較高的員工啟動留才管理機制。



# Introduction

## 資料描述

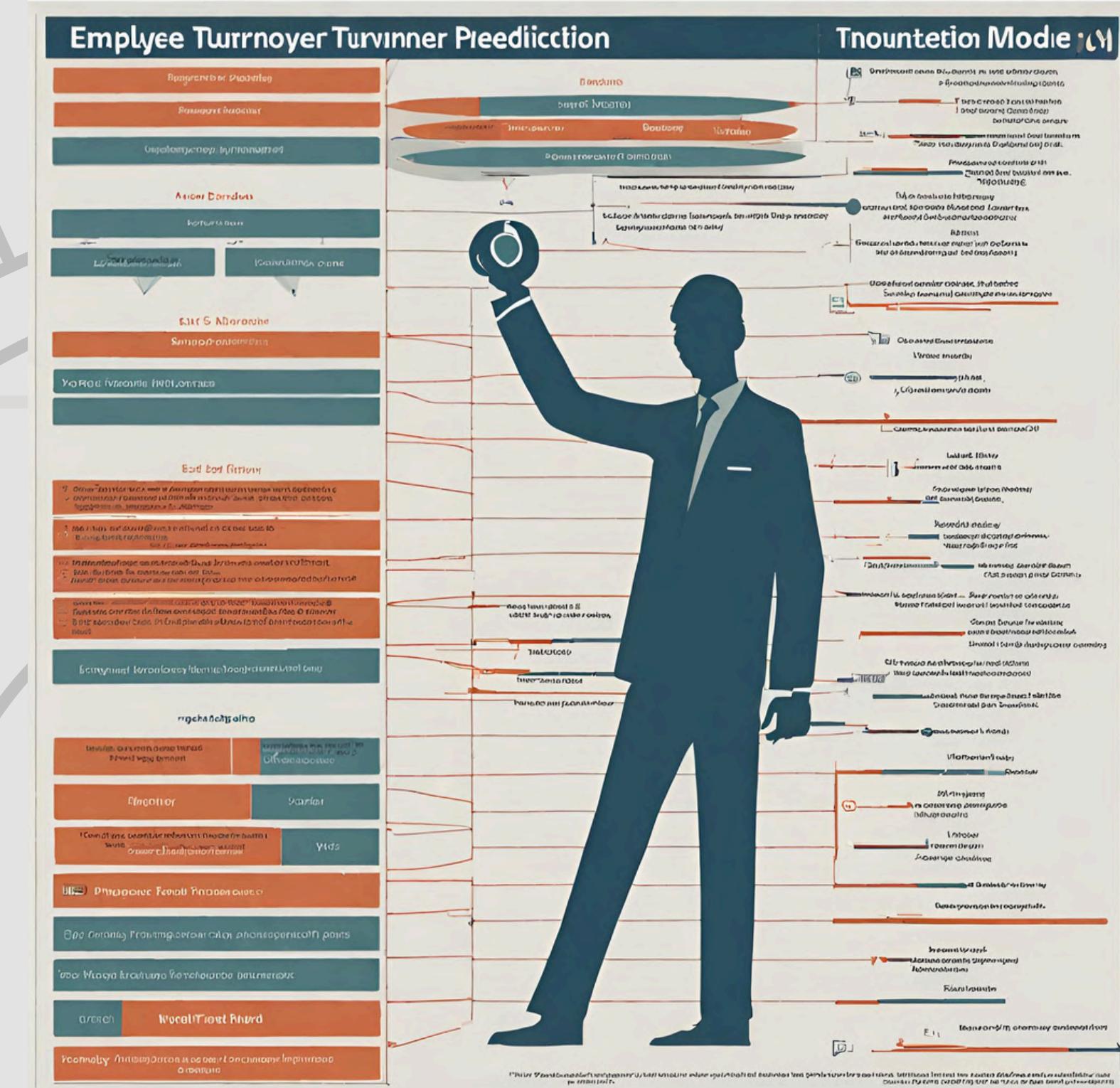
- 採用AIdea人工智慧共創平台提供的員工離職資料，訓練集包含14392筆資料，測試集包含3739筆資料，每筆資料皆有47個欄位。這些欄位特徵包括性別、職等、工作資歷等。其中，PerStatus欄位為預測目標，0代表未離職，1代表離職。
- 可能影響員工離職的因素，包括年齡層、績效、最高學歷、出差數、請假數等，這些因素將用於建立離職預測模型。

```
Index(['yyyy', 'PerNo', 'PerStatus', 'sex', '工作分類', '職等', '廠區代碼', '管理層級',
       '工作資歷1', '工作資歷2', '工作資歷3', '工作資歷4', '工作資歷5', '專案時數', '專案總數', '當前專案角色',
       '特殊專案佔比', '工作地點', '訓練時數A', '訓練時數B', '訓練時數C', '生產總額', '榮譽數', '是否升遷',
       '升遷速度', '近三月請假數A', '近一年請假數A', '近三月請假數B', '近一年請假數B', '出差數A', '出差數B',
       '出差集中度', '年度績效等級A', '年度績效等級B', '年度績效等級C', '年齡層級', '婚姻狀況', '年資層級A',
       '年資層級B', '年資層級C', '任職前工作平均年數', '最高學歷', '畢業學校類別', '畢業科系類別', '眷屬量',
       '通勤成本', '歸屬部門'],
      dtype='object')
```

# Method

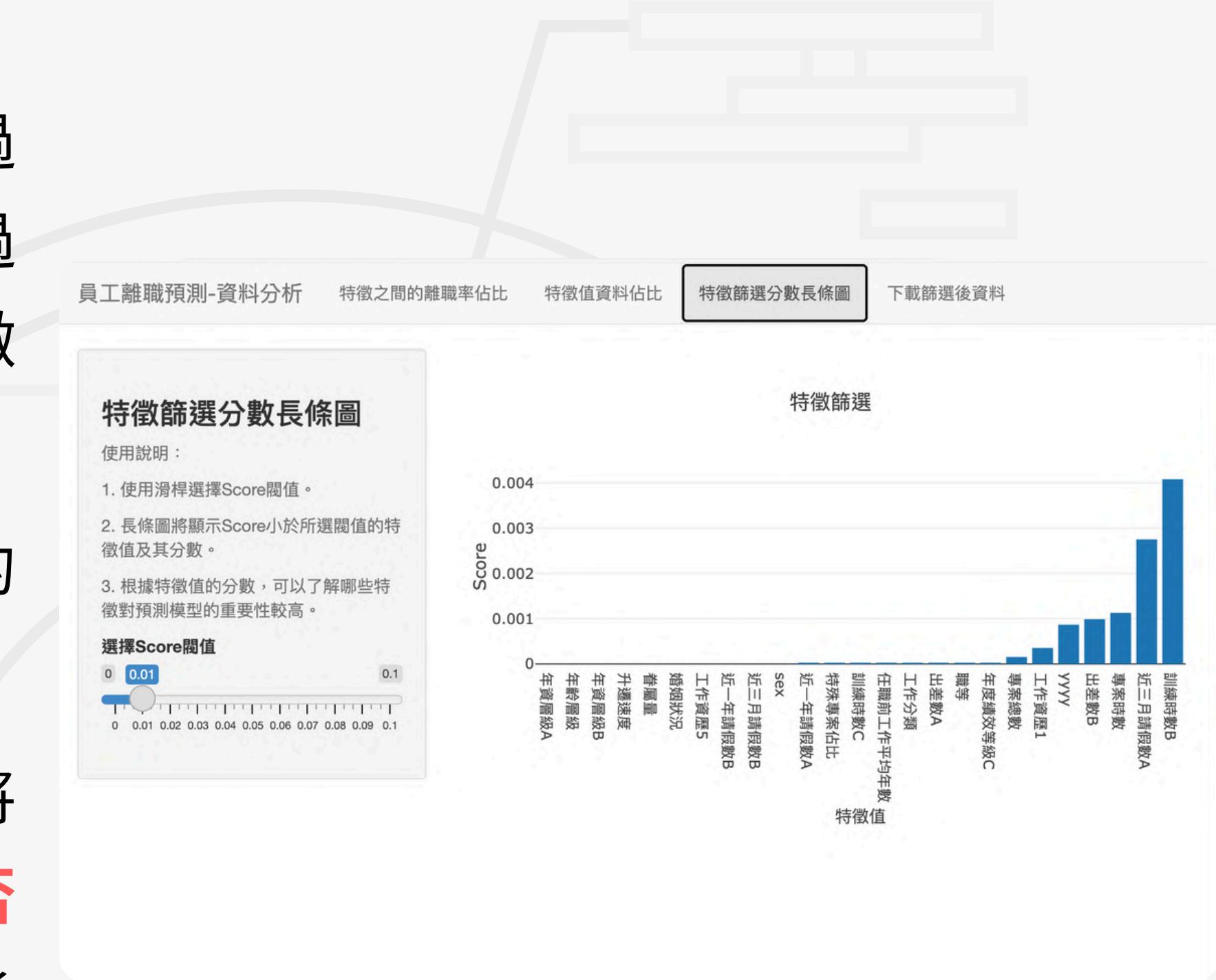
## Rshiny - 數據分析

## Python - 機器學習



# Rshiny - 資料分析

- 資料包含train.csv和test.csv，透過Rshiny來呈現一個網頁式介面，並透過Tabsets建立多種圖表來呈現不同特徵和離職之間的關係。
- 透過不同的圖表統計，了解造成離職的原因可能有哪些特徵。
- 同時會在Rshiny中處理資料，例如將**NA的欄位改為-1、計算出的特徵與是否離職的相關性較小**等等，接著將處理後的資料下載至本地端並投入模型訓練。

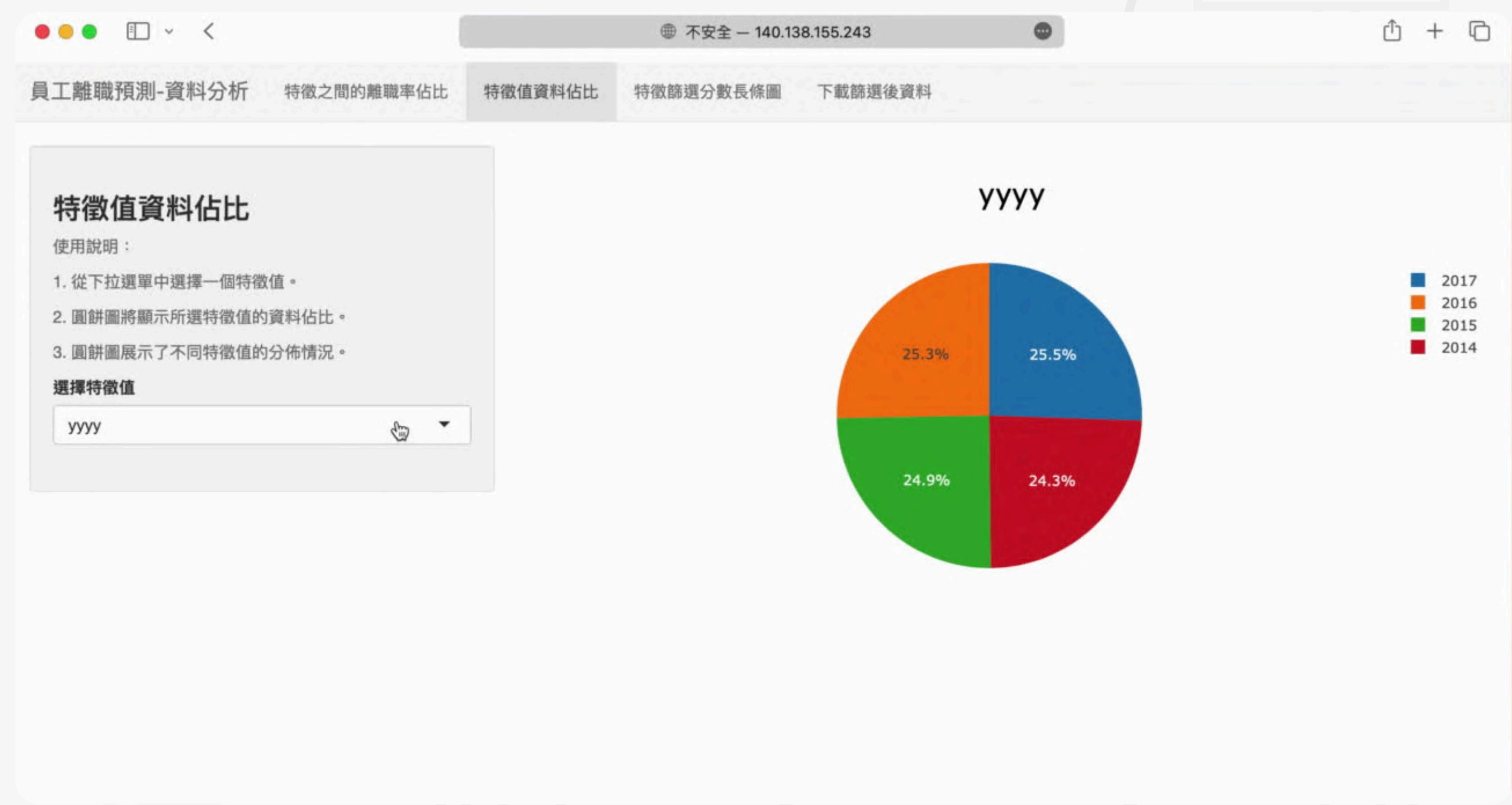


▲ Rshiny 介面

# Rshiny資料分析 - 特徵之間的離職率佔比



# Rshiny資料分析 - 特徵值資料佔比



# 使用 ANOVA 選擇特徵

## 變異數分析 (ANOVA)：

- 是一種統計方法，用於比較多個樣本均值之間的差異。
- 在特徵選擇中，ANOVA 可以幫助我們評估每個特徵與目標變數之間的關聯性，篩選出最重要的特徵。
- $p$  值越小，表示特徵越重要，可以篩選出與目標變數（員工離職）關聯性最強的特徵，助於進一步的數據分析和建模。
- 這些重要特徵可以用於構建更準確的預測模型，提高預測的準確性。

# Rshiny資料分析 - 特徵篩選分數長條圖



# Rshiny資料分析 - 下載篩選後資料

不安全 – 140.138.155.243

## 下載篩選後資料

使用說明：

1. 使用滑桿選擇Score閾值。
2. 從下拉選單中選擇所需的資料類型 (train.csv 或 test.csv)。
3. 根據特徵分數閾值篩選後的資料會顯示在表格中。
4. 選擇所需的特徵值，並從中選擇要包括在下載文件中的特徵。
5. 選擇下載文件的編碼格式 (UTF-8 或 Big5)。
6. 點擊下載按鈕以CSV格式下載篩選後的資料。

選擇Score閾值

0.01

選擇篩選後的特徵

年資層級A 年齡層級 年資層級B 升遷速度 善屬量  
婚姻狀況 工作資歷5 近一年請假數B  
近三月請假數B sex 近一年請假數A 特殊專案佔比  
訓練時數C 任職前工作平均年數 工作分類 出差數A  
職等 年度績效等級C 專案總數 工作資歷1 yyyy  
出差數B 專案時數 近三月請假數A 訓練時數B

選擇資料類型

PerNo	PerStatus	年資層級A	年齡層級	年資層級B	升遷速度	善屬量	婚姻狀況	工作資歷5	近一年請假數B	近三月請假數B	sex	近一年請假數A	特殊專案佔比	
1	1	0	2	6	1	1	0	1	0	0	0	1	5	3
2	1	0	2	7	2	1	2	1	0	0	0	1	6	1
3	1	0	2	7	2	1	2	1	0	0	0	1	7	1
4	1	0	2	7	2	2	2	1	0	0	0	1	5	5
5	3	0	5	9	5	6	2	1	0	0	0	0	6	1
6	3	0	6	10	6	6	2	1	0	0	0	0	6	3
7	3	0	6	10	6	7	2	1	0	0	0	0	7	3
8	3	0	6	10	6	7	2	1	0	0	0	0	7	3
9	6	0	6	11	4	6	1	1	0	0	0	1	6	1
10	6	0	6	11	5	7	1	1	0	0	0	1	6	0

Showing 1 to 10 of 14,392 entries

Previous 1 2 3 4 5 ... 1,440 Next

# Python - 機器學習

1. 由於會有資料不平衡的問題，因此會使用SMOTE來合成少數樣本。
2. 部分類別資料（如：工作分類）以**One hot encoding**處理。
3. 將處理後的資料投入模型訓練，投入的模型包括DNN、Decision Tree、XGBoost、Random Forest、OneClassSVM，**實測結果為DNN效果最好，在Aldea上取得第3名的成績，和第1名僅差距約0.002分**。
4. 模型訓練完成後，透過**SHAP**將模型的預測解釋分析成每個因子的貢獻，計算每個特徵的shapely value，來衡量該特徵對預測的貢獻度。

Public Leaderboard				Private Leaderboard				
檔名	上傳時間	評估結果	排名	排名	成績	上傳時間	次數	
DNN_predicted2.csv	2024-06-20 00:30:23	0.2731884	3/604	1	weisen91617	0.2753099	2022/01/18 19:56:09	15
Chih Hung Han				2	fanfan	0.2747584	2021/06/21 04:37:30	6
				3	Chih Hung Han	0.2731884	2024/06/20 00:30:23	28

# Python - Data Preprocess : SMOTE

離職員工和未離職員工的數量存在較大差異：為什麼數據不平衡是一個問題？

- **模型偏倚：**  
在不平衡數據集上訓練的模型可能會偏向多數類別（未離職），即使模型在多數類別上的準確率很高，但在**少數類別（離職）上的預測性能可能很差**
- **性能指標失真：**  
標準的性能指標（如準確率）在不平衡數據集上可能會誤導。例如90%的員工未離職，那全預測未離職也能達到90%的準確率

# Python - Data Preprocess : SMOTE

Synthetic Minority Oversampling Technique **合成少數過採樣**技術，概念是對少數類的樣本進行分析，並人工合成新樣本添加到數據集中，通過在少數類別樣本之間插值生成新的合成樣本，增加少數類別的樣本數量，在平衡數據集上訓練的模型可以**更好地學習少數類別的特徵**，提高模型對少數類別的預測性能

## 可能遇到的問題：

- 生成的合成樣本不符合數據的真實分佈、過擬合訓練數據
- 如果數據集中存在異常值並被用來生成新的合成樣本會影響模型性能

# Python - Models and Shap

1. 使用模型：Decision Tree、DNN、XGBoost 、Random Forest、oneclass SVM

2. SHAP分析：

- SHAP (SHapley Additive exPlanations) 一種解釋機器學習模型的方法，SHAP 值是每個特徵對模型預測結果的貢獻

3. SHAP 值：

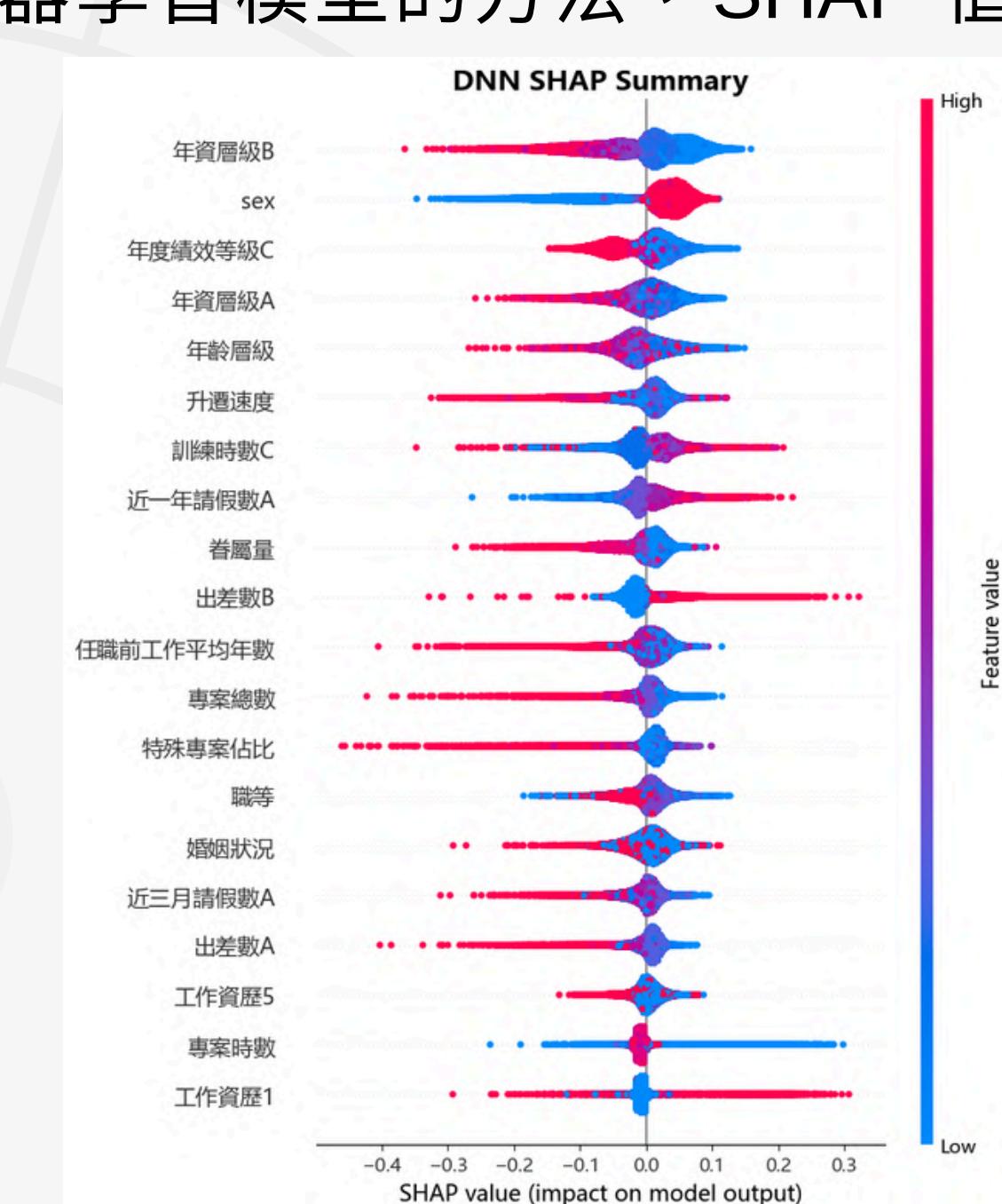
- 表示特徵對預測結果的影響
- SHAP 值  $> 0$ ：增加離職可能性（輸出1）
- SHAP 值  $< 0$ ：減少離職可能性（輸出0）

4. 點的顏色：

- 紅色：特徵值高、藍色：特徵值低

5. 左右擴散程度：

- 擴散範圍越大，表示該特徵對預測結果的影響越大



# Python - Models and Shap

## 如何解讀 SHAP 圖？

### 1. 特徵重要性：

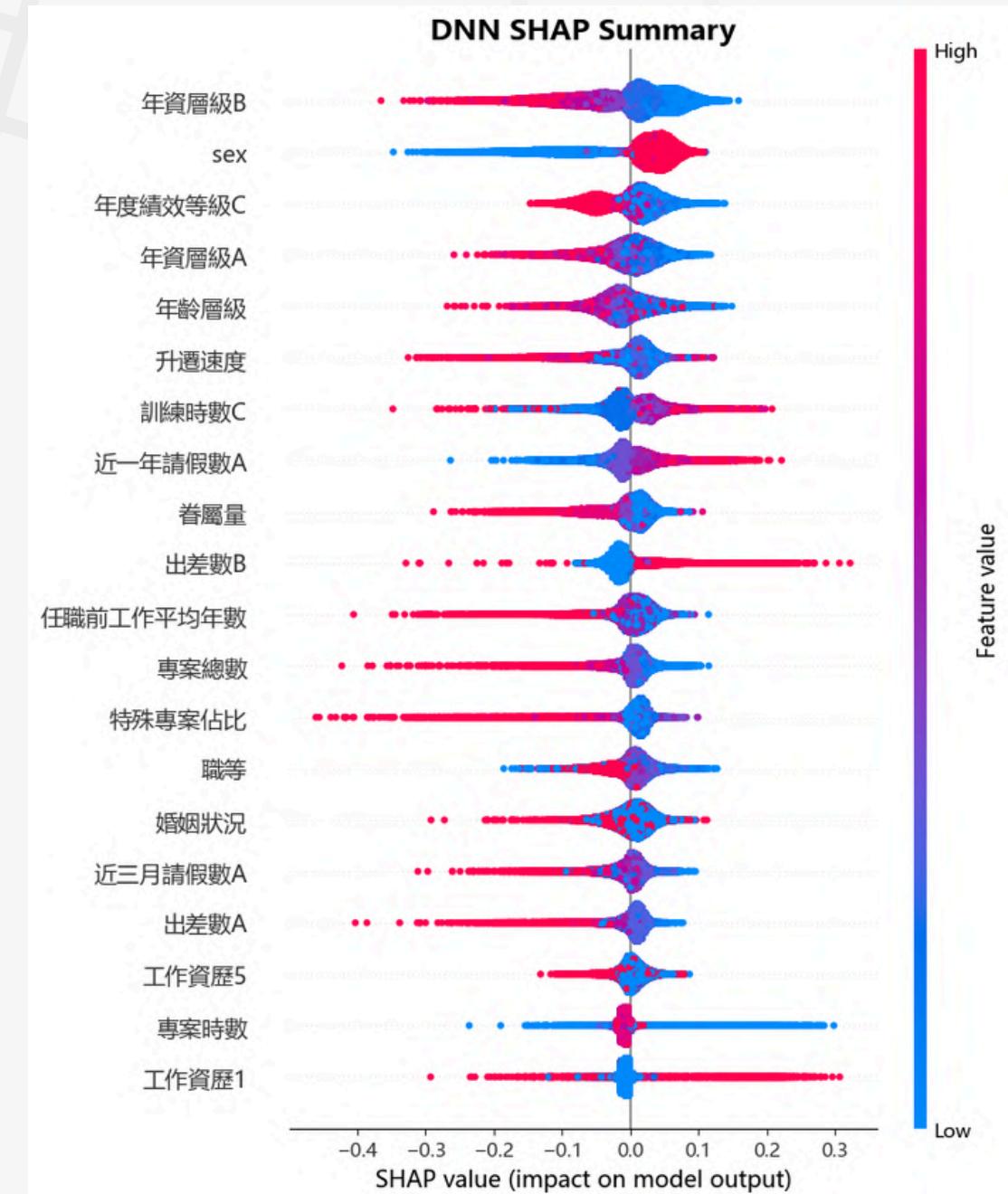
- 特徵排列順序代表重要性，越上面的特徵對預測結果影響越大

### 2. 正 SHAP 值（右側）：

- 增加離職的可能性
- 紅色點在右側表示高特徵值，例如：高工齡增加離職可能性

### 3. 負 SHAP 值（左側）：

- 減少離職的可能性
- 紅色點在左側表示高特徵值，例如：高薪水減少離職可能性



# Results

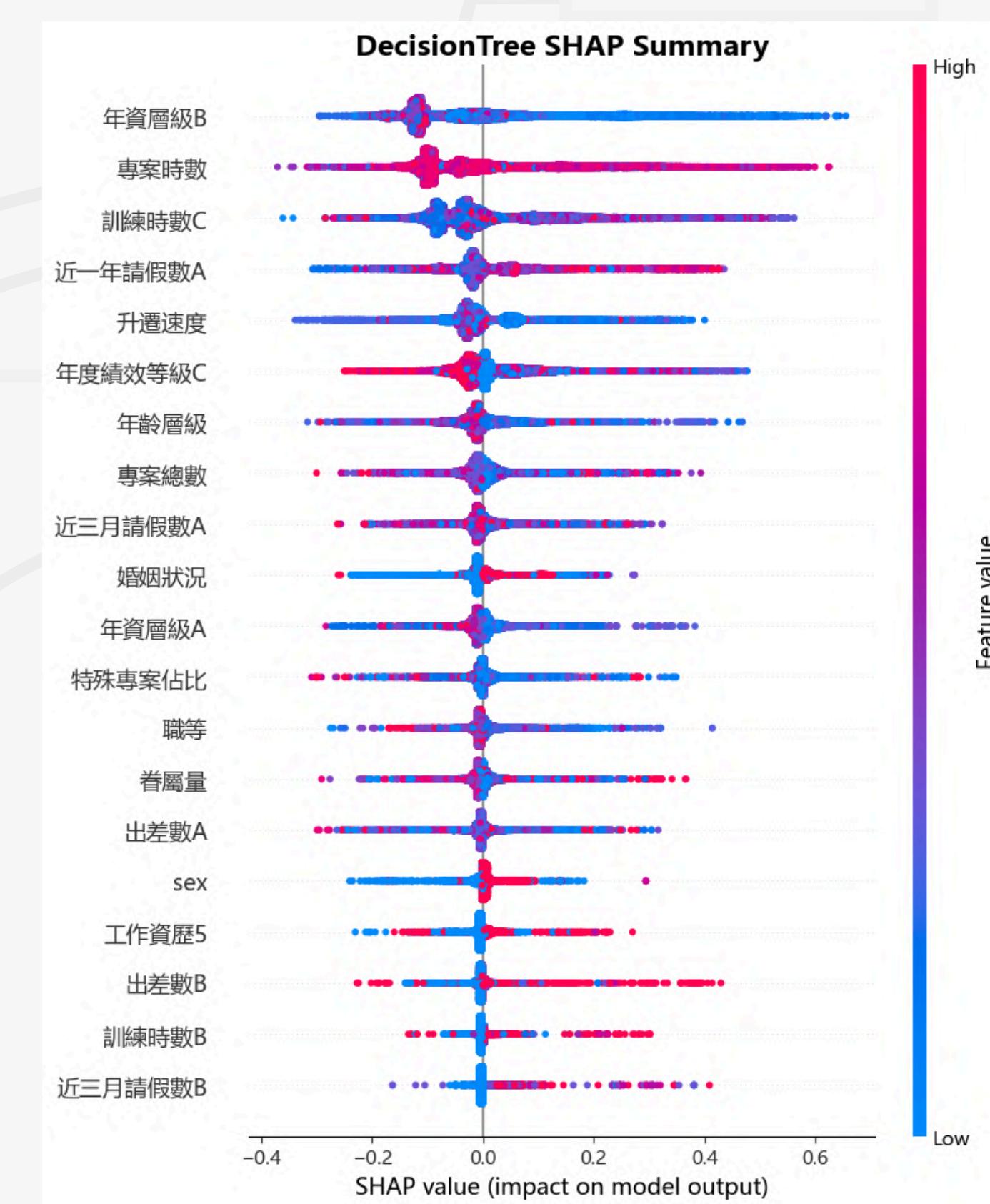
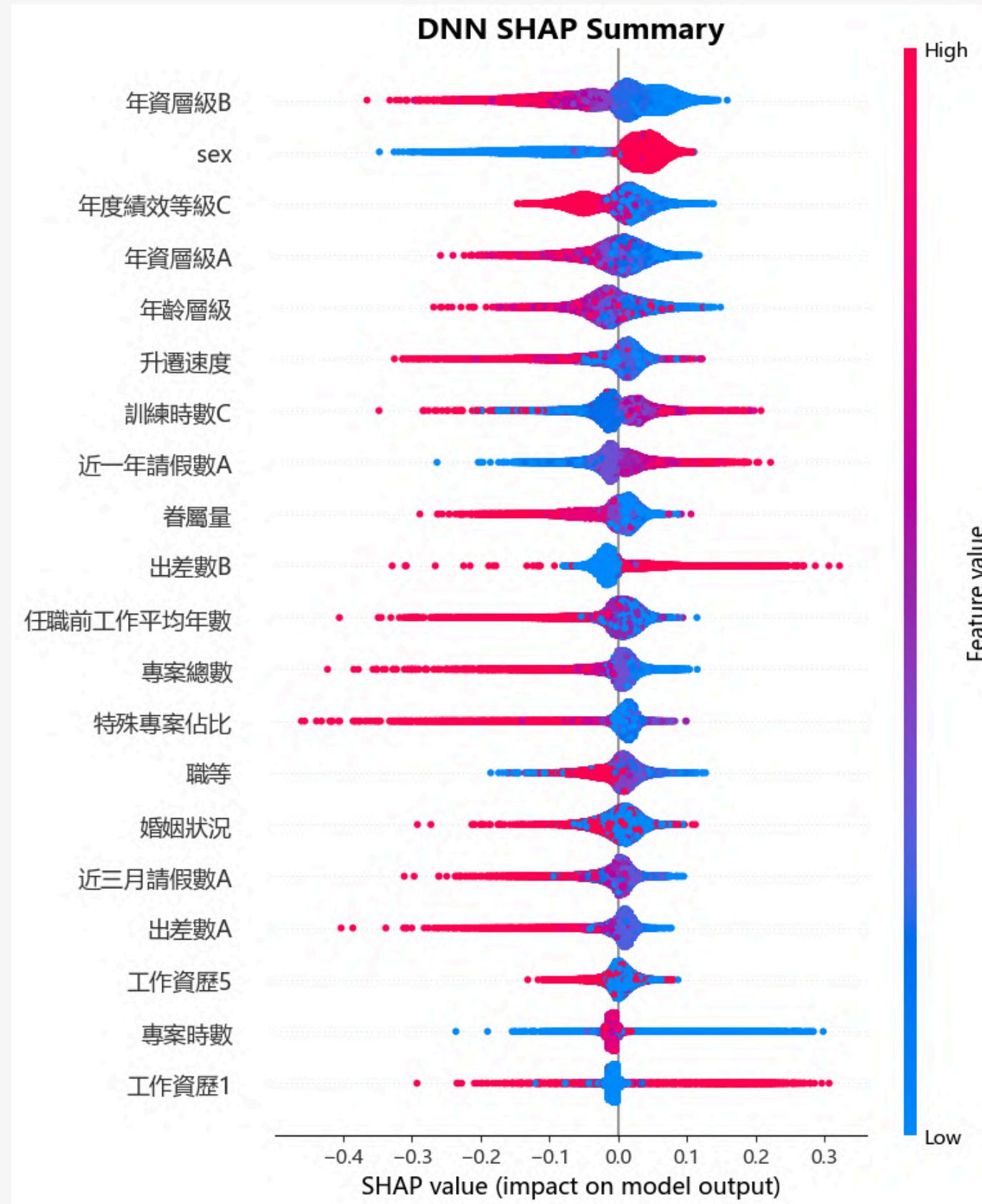


# Predicted Results

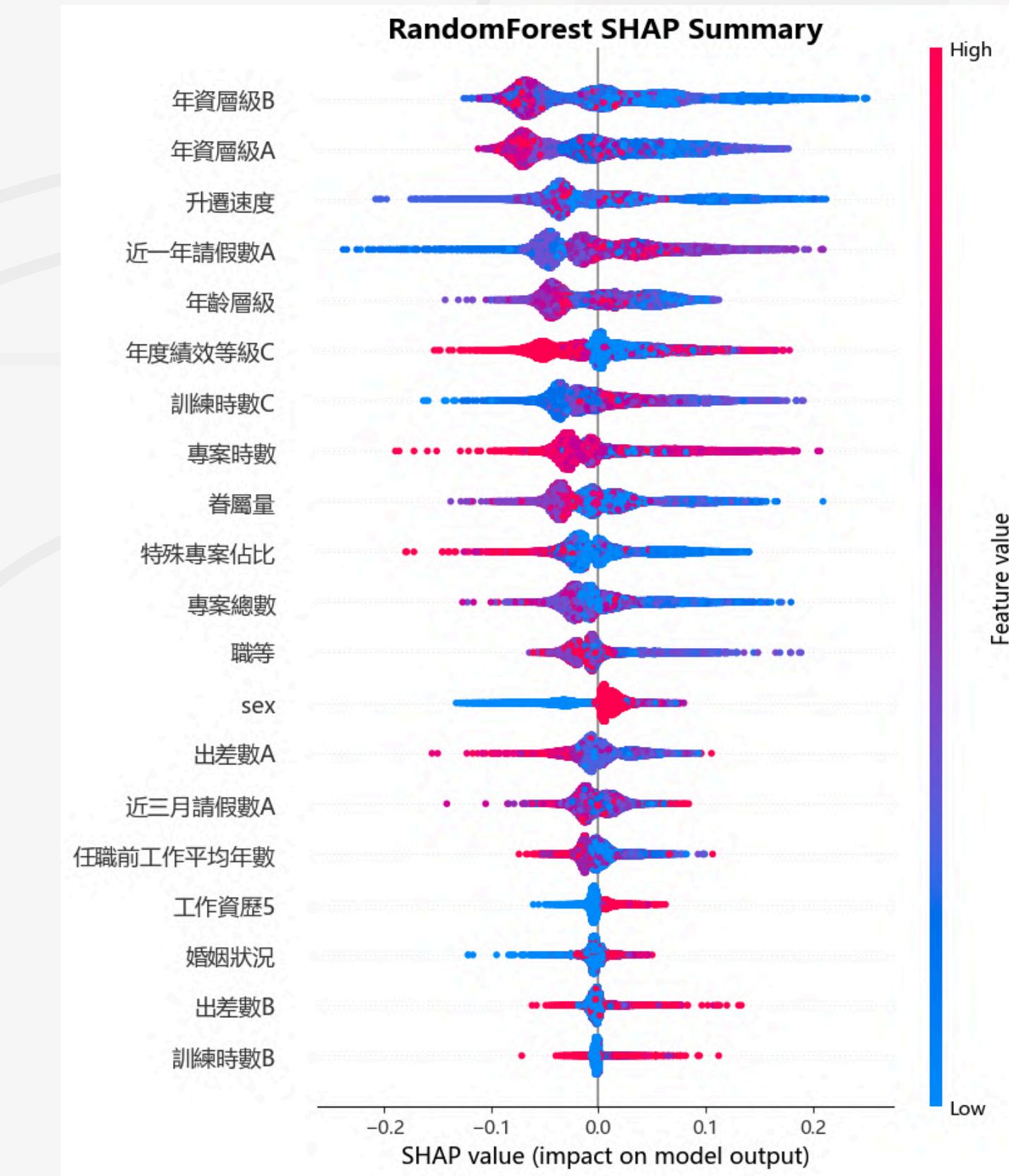
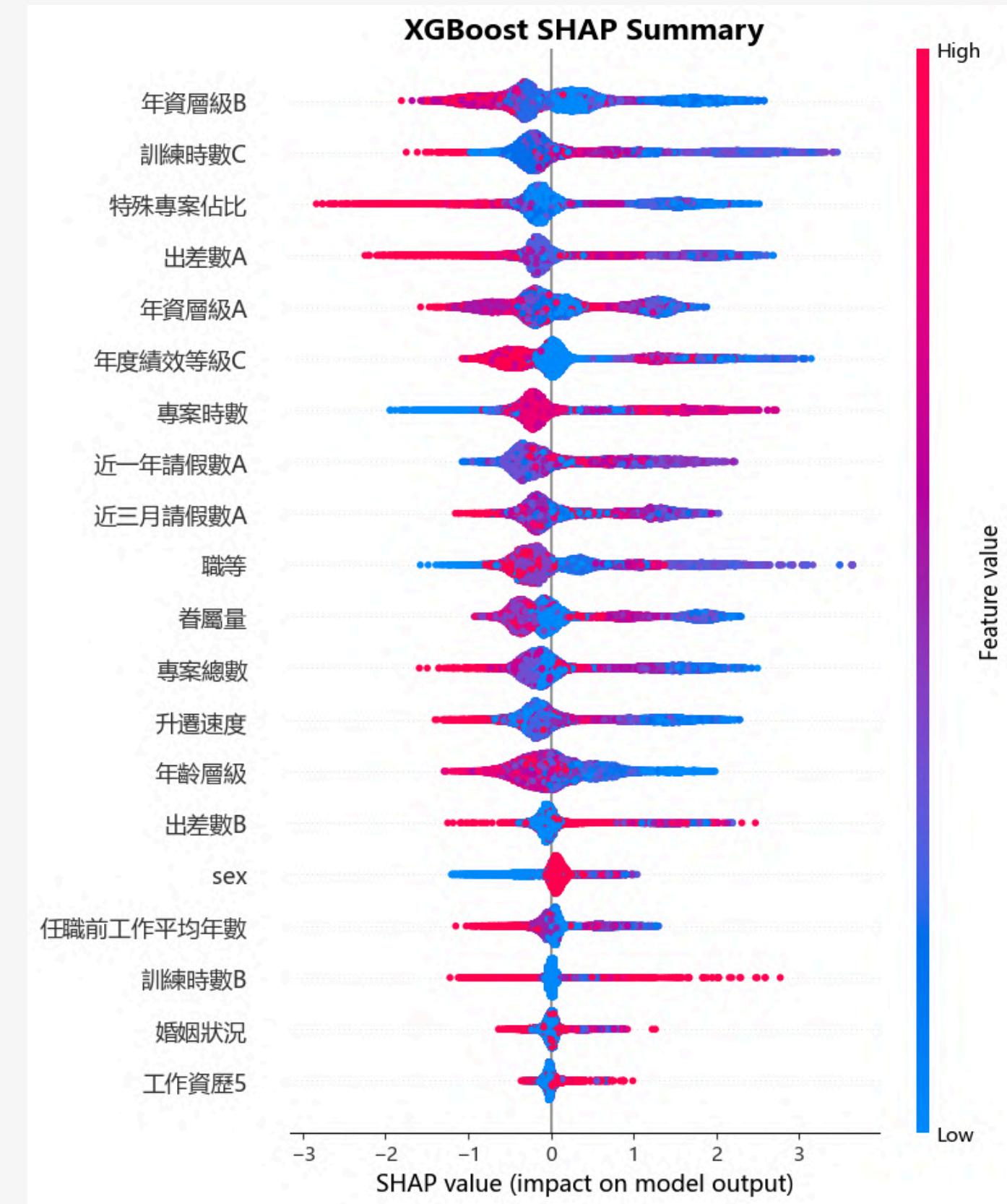
$$F_{\beta} = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

Model	F 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

# Python - Shap Summary : DNN、Decision Tree



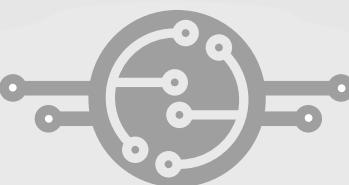
# Python - Shap Summary : XGBoost、Random Forest



# Conclusion

- ANOVA 篩選特徵
- DNN 的 F1 Score 和排名最高，這可能是因為它能夠捕捉數據的複雜模式
- XGBoost 雖然在很多比賽表現很好，但在排名較差，這可能是因為需要進行更精細的 **參數調整**，才可以發揮他的能力
- Random Forest 和 Decision Tree 可能是因為數據的特徵比較複雜，需要更多的 **調整和修剪 (pruning)**
- OneClass SVM 的表現最差，可能是因為它通常用於 **異常檢測** 非二元分類問題。
- SMOTE 影響

# 大數據資料分析期末專題



# THANKS