EMPLOYEES ATTRITION OF IBM

統計碩二 林子涵、統計大三 邱士芳、資管大三 黃茂勛
廣電大三 賴冠儒、英文大三 晏煒翔、資科碩一 管漢程

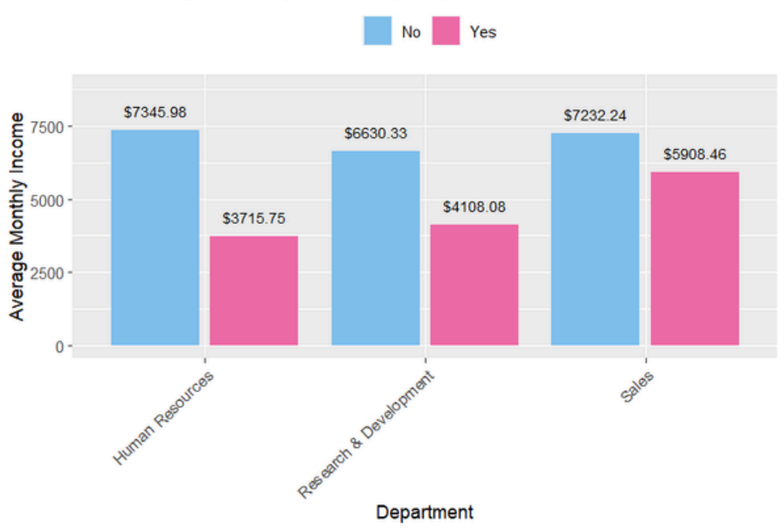
2024 Data Science Group 1

INTRODUCTION

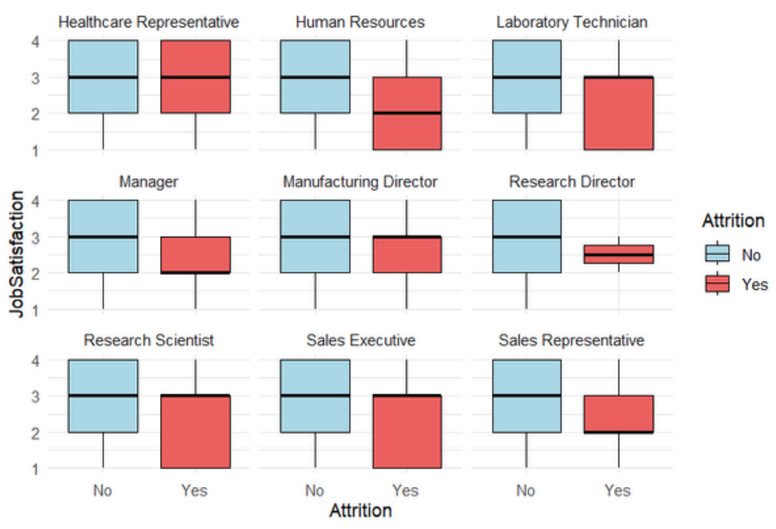
我們的小組專案旨在分析IBM員工的基本資料與員工流失率（employee attrition）之間的關係。通過深入研究員工的人口統計數據，例如年齡、性別、教育背景、工作年限等，我們希望找出可能影響員工離職的關鍵因素。本專案將結合統計分析和數據視覺化技術，提供實證資料來支持管理決策。

EDA

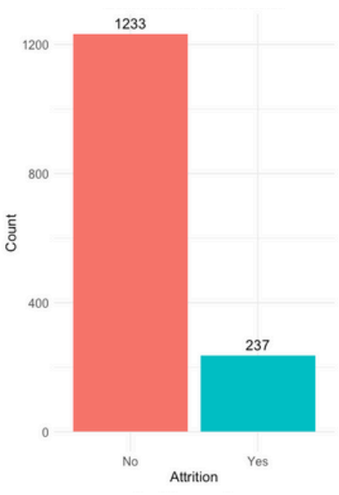
Average Monthly Income by Department and Attrition Status



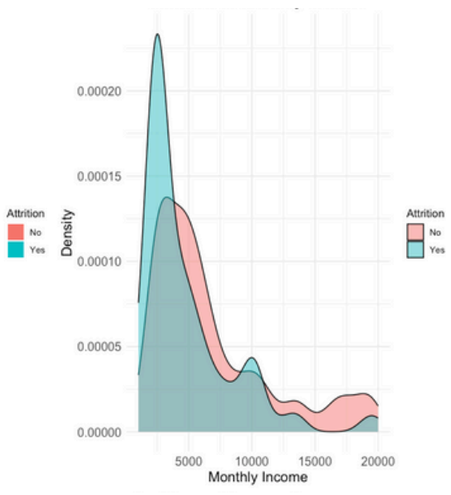
Average Boxplot of Job Satisfaction by Attrition and Job Role



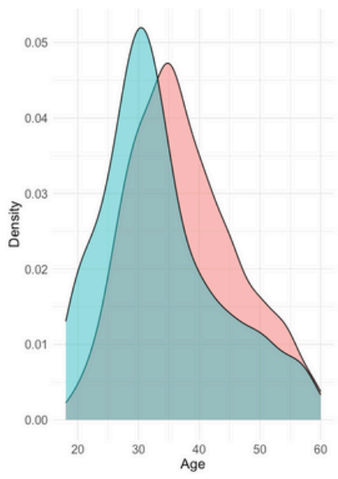
Distribution of Attrition



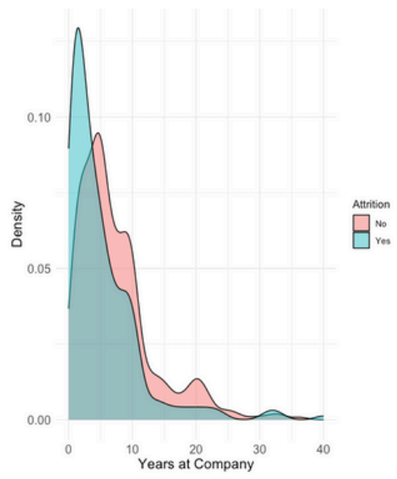
Attrition vs Monthly Income



Attrition vs Age



Attrition vs Years at Company



METHOD&MODEL

null model

- Accuracy：0.8386
- Kappa：0
- Sensitivity : 1.0000
- Specificity : 0.0000
- Pos Pred Value : 0.8386

雖然準確率高達83.86%，但Kappa係數為0，意味著模型的預測能力與隨機猜測無異。模型能100%預測未離職員工，但無法識別離職員工，這說明需引入更多變數進行進一步分析

Random Forest

為避免過度擬合，我們設定k = 5進行交叉驗證

fold	accuracy	precision	recall	f1
1	0.8738	0.8731	0.9942	0.9297
2	0.8592	0.8557	1.0000	0.9223
3	0.8883	0.9031	0.9779	0.9390
4	0.8300	0.8300	0.9940	0.9046
5	0.8495	0.8571	0.9825	0.9155
mean	0.8602	0.8638	0.9897	0.9222

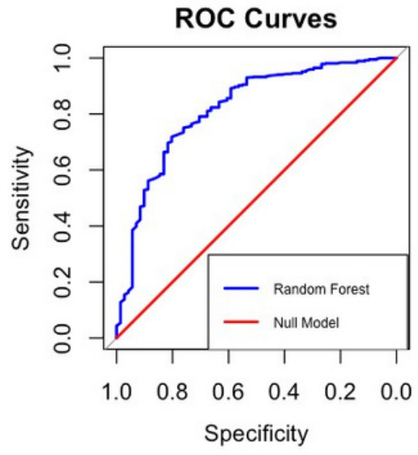
預測結果

Prediction	No	Yes
No	362	56
Yes	7	15

- Accuracy: 0.8568
- Kappa: 0.2666
- Sensitivity: 0.9810
- Specificity: 0.2113
- Pos Pred Value: 0.8660
- Neg Pred Value: 0.6818
- Balanced Accuracy: 0.5961

AUC

- null model : 0.5
- Random Forest : 0.8185



重要變數

使用 importance 函數分析各個變量的重要程度，數值越大代表越重要。

- 月收入：27.2079
- 年齡：21.5346
- 是否加班：18.6995
- 離家距離：16.6699
- 總工作年限：16.7378
- 工資增幅：13.2598
- 待過公司數：12.6056
- 工作年數：11.7368
- 工作職位：10.9280
- ……

重要變數依序為：月收入、年齡、是否加班、離家距離、總工作年限

CONCLUSION

我們的研究以 IBM 員工離職率為主題，利用 null model 和 Random Forest 模型進行分析。結果顯示，儘管 null model 具有高準確率，但 Kappa 係數為0，需要更多變數進行深入分析。隨後使用 Random Forest 模型，準確率為0.8601，並且 Kappa 系數提高至0.2666，顯示模型的預測能力有所提升。