公司員工離職預測

統計碩一 107354012
陳冠廷

目錄

PART 1

資料介紹

- 研究目的
- 資料來源
- 變數解釋
- 資料探索

研究目的

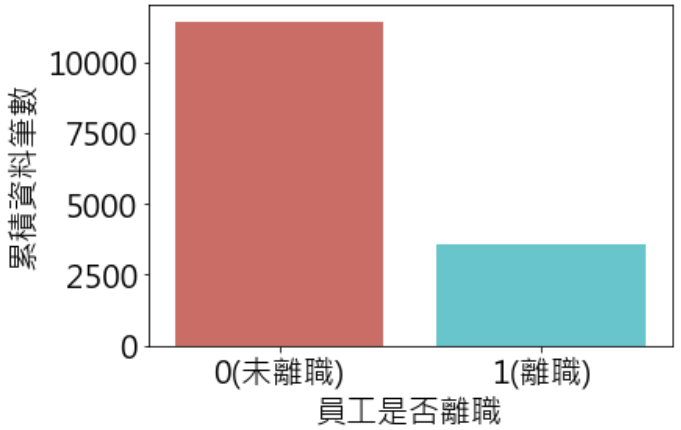
員工一直是影響公司營運的重要因素之一，因此如何解決優秀員工流失的問題，是世界許多公司正面臨的挑戰。

這次我透過對員工離職數據集的分析，嘗試用 Logistic Regression 來預測有很大機率將要離職的員工及其背後影響員工離職的關鍵（員工滿意度、薪資水平、平均每月工時等），在了解對離職員工產生重大影響的因素後，便可以建議公司採取適當的措施來改善這些因素以留住公司人才。

資料來源

資料介紹

- 原始出處: kaggle
- 資料網址:https://www.kaggle.com/ludobenistant/hr-analytics
- 資料提供者:Ludovic Bénistant



- 資料筆數:14, 999筆員工資料(離職:未離職=23. 8:76. 2)，10個變數(9個解釋變數，一個反應變數)

	left	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years	position	salary
0	1	0.38	0.53	2	157	3	0	0	sales	low
1	1	0.80	0.86	5	262	6	0	0	sales	medium
2	1	0.11	0.88	7	272	4	0	0	sales	medium
3	1	0.72	0.87	5	223	5	0	0	sales	low
4	1	0.37	0.52	2	159	3	0	0	sales	low
5	1	0.41	0.50	2	153	3	0	0	sales	low
6	1	0.10	0.77	6	247	4	0	0	sales	low
7	1	0.92	0.85	5	259	5	0	0	sales	low

變數解釋

反應變數

Left

是否離職
值為0或1
0代表未離職
1代表已離職

解釋變數

satisfaction_level

員工滿意度
Numeric:0~1

average_montly_hours

平均每月工時
Integer:96~310

number_project

參與過的專案數
Integer:2~7

last_evaluation

績效評估
Numeric:0~1

time_spend_company

進入公司的年數
Integer:2~10

position

在公司所屬部門
類別型變數
共10種職位

Work_accident

是否有過工作意外
值為0或1
0代表未發生
1代表曾發生

promotion_last_5years

五年內是否升職
值為0或1
0代表未升職
1代表已升職

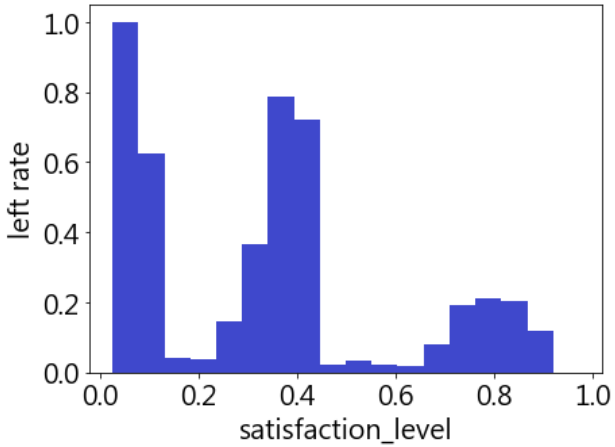
salary

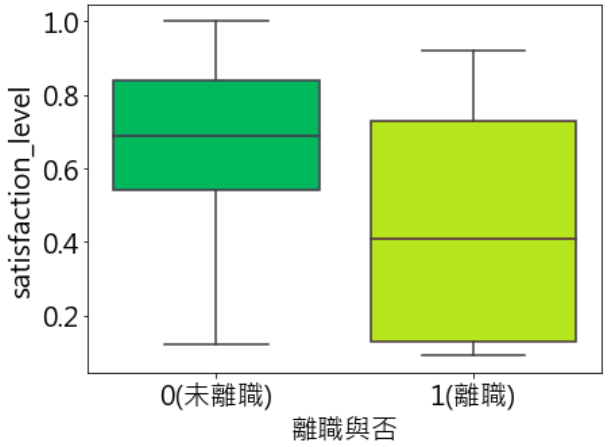
薪資水平
類別型變數
共分為
High,Medium,Low
三個水準

資料探索

satisfaction_level

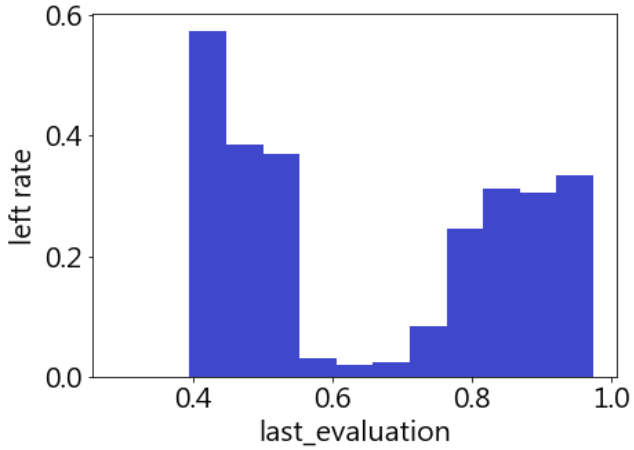
整體來看，員工滿意度越高，離職率越低

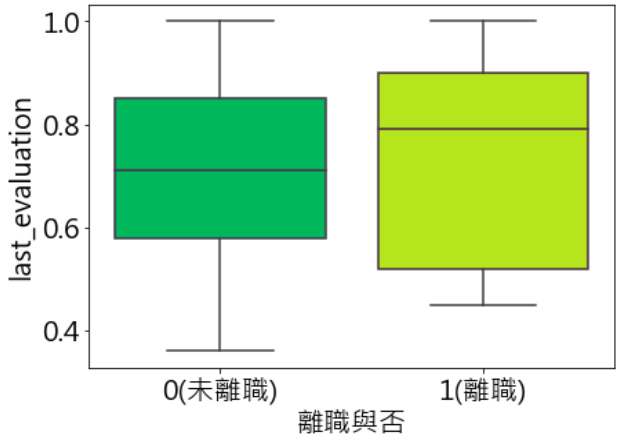




last_evaluation

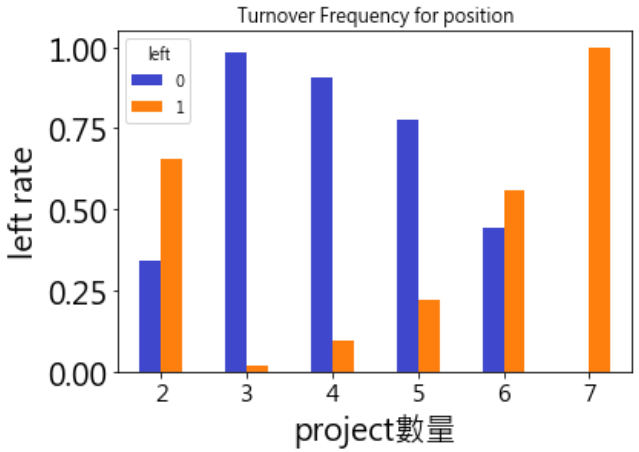
離職率與績效評估看似沒有關聯

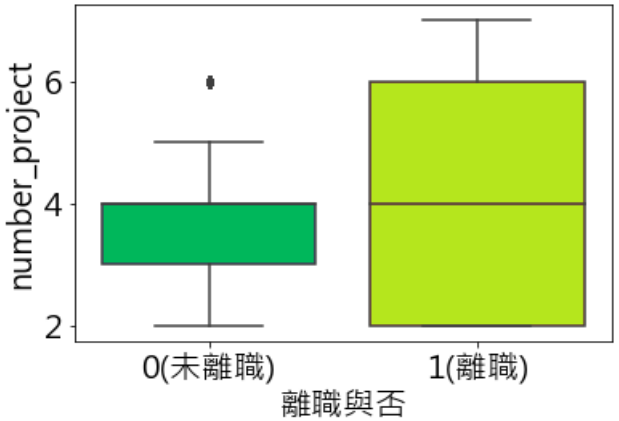




number_project

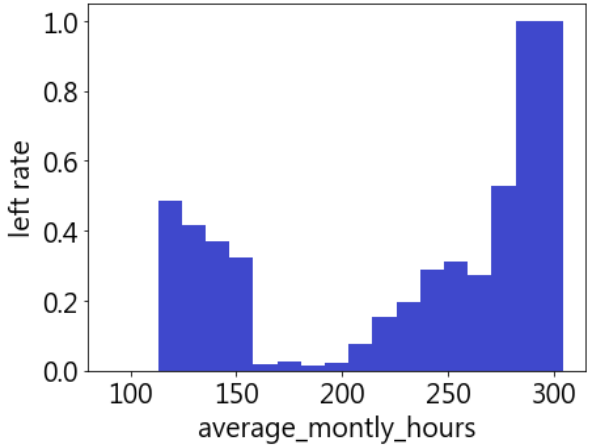
專案數越大，離職率越大

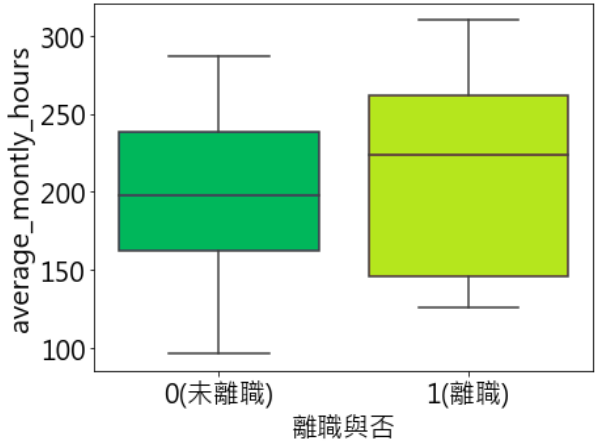




average_montly_hours

每月平均工時越多，離職率越大



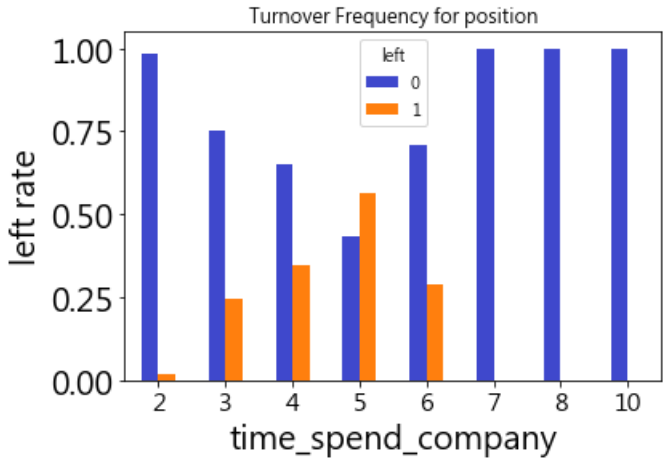


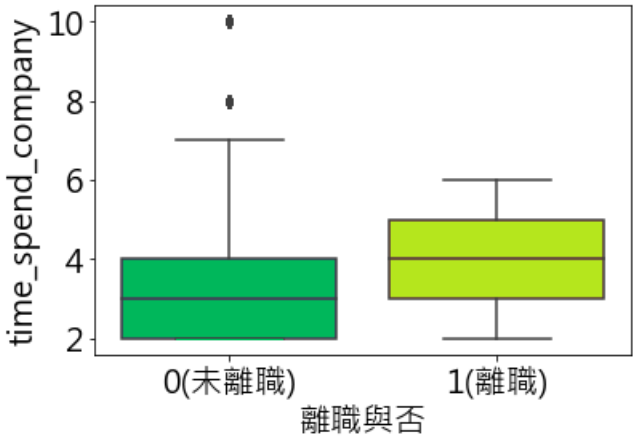
資料探索

資料介紹

time_spend_company

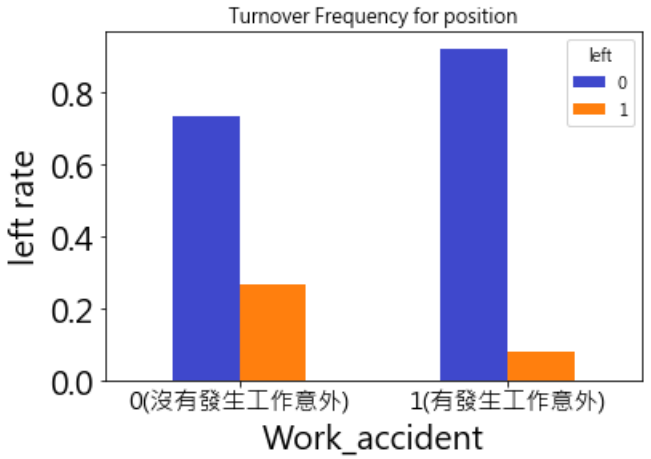
工作五年內，離職率隨工作年數增加





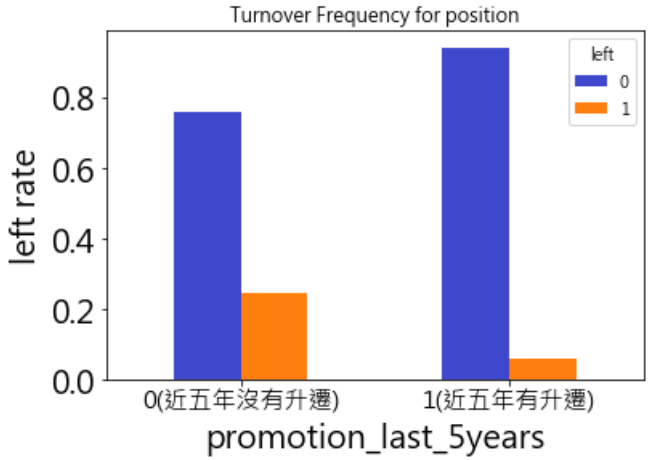
Work_accident

有無發生意外與離職率沒有太大差異



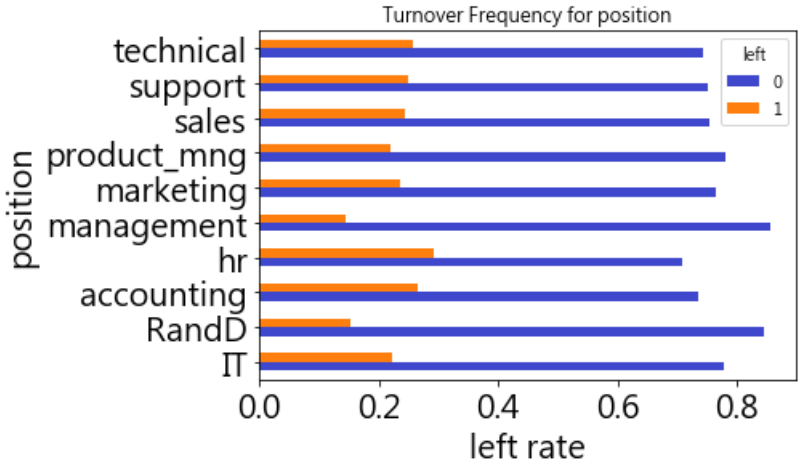
promotion_last_5years

有無升遷與離職率沒有太大差異



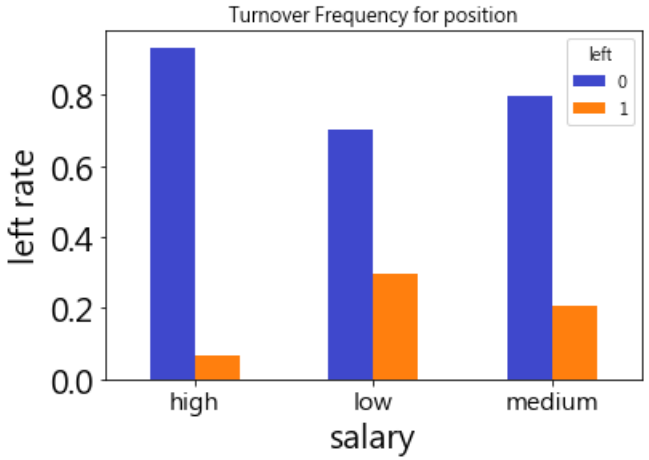
Position

所在部門與離職率沒有太大差異



Salary

薪水水平為高者較不易離職



PART 2

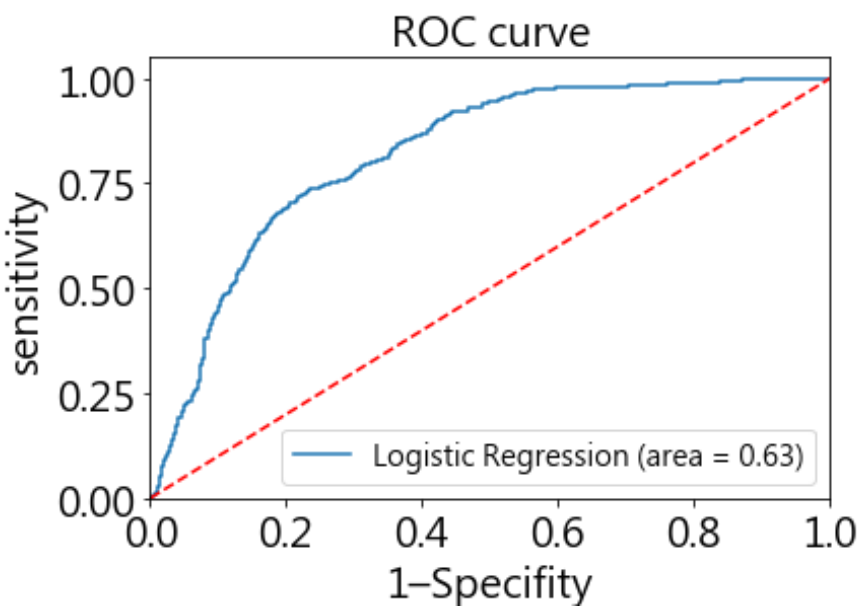
資料整理

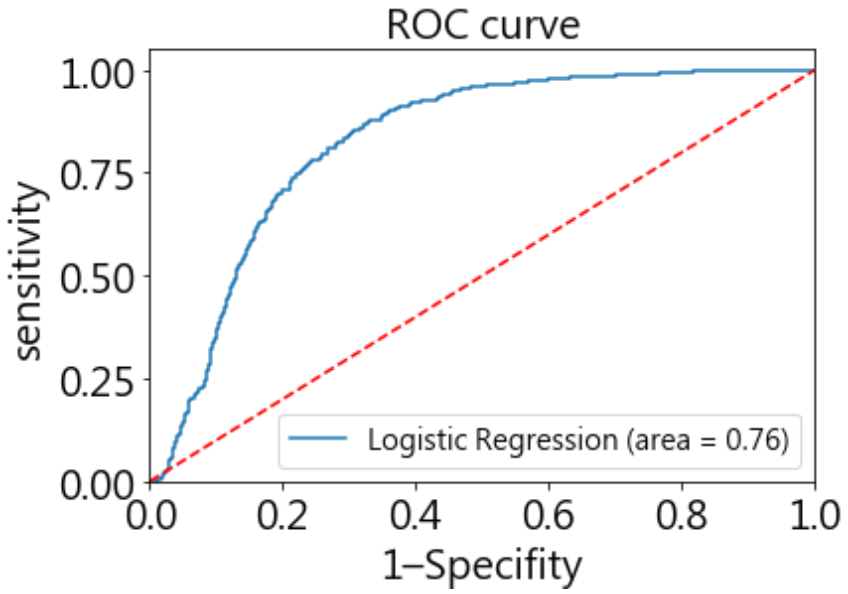
- 資料不平衡處理
- 資料集拆分
- 變數處理

資料不平衡處理

此份資料為不平衡資料(離職:未離職=23. 8:76. 2)

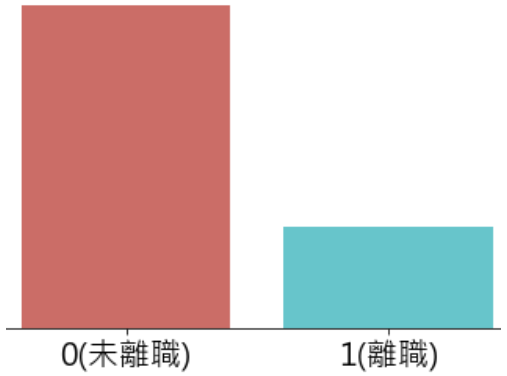
- 採取SMOTE方法(Synthetic Minority Over-Sampling Technique)模擬生成樣本，
  從離職樣本中生成樣本使得離職與未離職樣本數相同
- SMOTE運作概述: 1.對於每一個少數樣本s，以歐式距離計算該樣本與其他少數樣本的距離
  2.以此距離找出與該樣本s最近的K個樣本
  3.從K個樣本中選取M個樣本
  4.將該樣本s在M個樣本中的每一個樣本r_i透過$S_new=\lambda s+(1-\lambda)\ r_i$生成新樣本S_new,
  其中λ為0~1的任意數
- 採用SMOTE方法後，AUC獲得改善(0. 63➔0. 76)



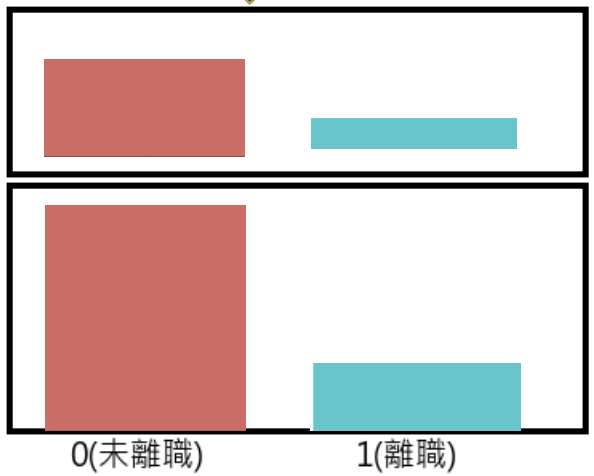


資料集拆分

資料整理



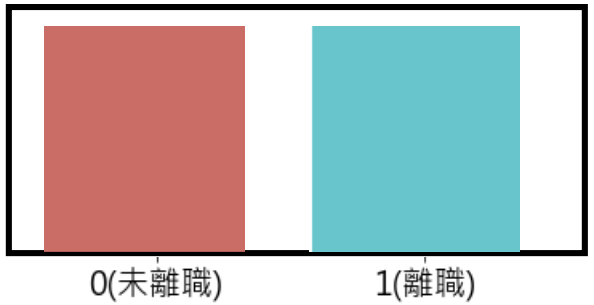
分別抓出離職以及未離職之員工資料



測試集

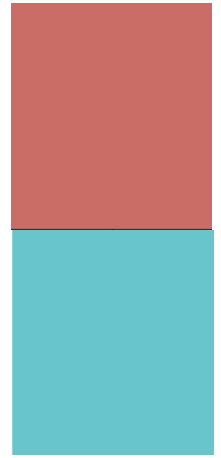
訓練集

分別將兩種資料按3:7拆分測試集及訓練集



訓練集

將離職以及未離職之員工訓練集進行SMOTE使兩種資料樣本數相同



模型訓練集

合併離職以及未離職之員工SMOTE後之訓練集作為模型之訓練集

模型測試集

合併離職以及未離職之員工之測試集作為模型訓練之測試集
(維持原數據集中離職與未離職比)

變數處理

- 將數值型變數歸一化，使變數範圍縮小到0~1之間，避免因數值大小使模型係數差異過大
- 類別型變數轉換:
  1.position欄位中分為IT,RandD,accounting,hr,management,marketing,producr_mng, sales,support,technical共10種，因此轉換為9個dummy variables
  2.salary欄位中分為High,Medium,Low共三種，因此轉換為2個dummy variables
  3.轉換後變數總數為18個
- 增加交互作用項:
  1.將類別型變數轉換後的變數兩兩相乘做一個交互作用項，增加C_2^{18} =153個變數
  2. 轉換後變數總數為171個
- 增加高次項:
  1. 將原數據集中的5個連續變數satisfaction_level, last_evaluation, average_montly_hours, number_project,time_spend_company增加二次項及三次項，共增加10個變數
  2.轉換後變數總數為181個

PART 3

模型配置

- 模型處理步驟
- 使用Lasso
- 使用逐步迴歸
- 確定最終模型

模型處理步驟

模型配置

拆分資料集

將Lasso篩選後的變數放入Logistic Regression 模型裡並去掉不顯著的變數

評估最終模型

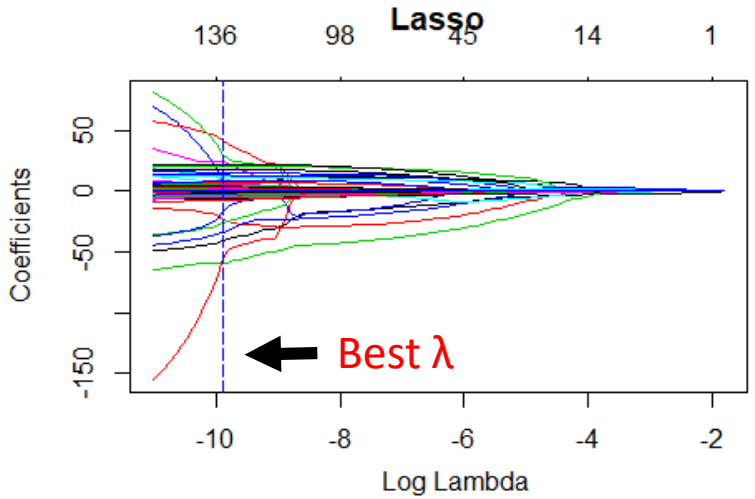
使用Lasso篩選變數

將剩餘變數分別帶入向前、向後、雙向逐步迴歸中，選擇AIC最低的作為最終模型

使用Lasso

模型配置

- 最佳λ為$5.05202*10^{-5}$
- 使用Lasso後變數由181個變為130個
- 將剩餘130個變數帶入Logistic Regression 模型裡並刪除P-value>0.01的變數
- 刪除後剩餘32個變數



Lasso變數收斂情形(列出部分)

	Estimate
(Intercept)	27.20006492
satisfaction_level	-40.95892796
last_evaluation	-23.65009906
number_project	-59.27808348
average_montly_hours	8.66861291
promotion_last_5years	.
Work_accident	-1.83021398

刪除P-value>0.01的變數(列出部分)

	Estimate	Pr(>|z|)
~~(Intercept)~~	~~7.118e-01~~	~~8.298828e-01~~
satisfaction_level	-55.20267092	3.022989e-53
~~last_evaluation~~	~~-12.19882159~~	~~3.928042e-02~~
number_project	-74.95249596	2.257499e-28
average_montly_hours	148.20130101	8.639483e-25
time_spend_company	20.46260773	2.272918e-05
~~Work_accident~~	~~-13.53175921~~	~~9.466631e-01~~

剩餘變數32個

模型配置

原始變數4個，高次項變數9個 ，交互作用項19個

[1] "satisfaction_level"
[2] "number_project"
[3] "average_montly_hours"
[4] "time_spend_company"
[5] "satisfaction_level^2"
[6] "satisfaction_level^3"
[7] "last_evaluation^2"
[8] "last_evaluation^3"
[9] "number_project^2"
[10] "average_montly_hours^2"
[11] "average_montly_hours^3"
[12] "time_spend_company^2"
[13] "time_spend_company^3"
[14] "satisfaction_level*last_evaluation"
[15] "satisfaction_level*average_montly_hours"
[16] "satisfaction_level*time_spend_company"
[17] "last_evaluation*number_project"
[18] "last_evaluation*average_montly_hours"
[19] "last_evaluation*time_spend_company"
[20] "last_evaluation*position_sales"
[21] "last_evaluation*salary_low"
[22] "last_evaluation*salary_medium"
[23] "number_project*average_montly_hours"
[24] "number_project*time_spend_company"
[25] "number_project*position_accounting"
[26] "average_montly_hours*time_spend_company"
[27] "time_spend_company*position_IT"
[28] "time_spend_company*position_management"
[29] "time_spend_company*position_sales"
[30] "time_spend_company*salary_medium"
[31] "position_management*salary_low"
[32] "position_management*salary_medium"

使用逐步迴歸

- 使用向前逐步迴歸，分析後之變數剩29個，減少average_montly_hours^2, average_montly_hours^3, number_project*position_accounting，AIC為6476.3
- 使用向後逐步迴歸，分析後之變數剩31個，減少number_project*position_accounting，AIC為6252.2
- 使用雙向逐步迴歸，分析後之變數剩29個，減少average_montly_hours^2, average_montly_hours^3, number_project*position_accounting，與向前逐步迴歸相同，AIC為6476.3
- 選擇AIC最低之向後逐步迴歸所得模型，作為最終模型

確立最終模型

模型配置

$\text{Logit}(\widehat{\text{left}}) =$

-8.38-49.14*satisfaction_level-57.13 *number_project +159.28*average_montly_hours
+31.86*time_spend_company + 54.04*satisfaction_level^2 -35.72* satisfaction_level^3
-58.6*last_evaluation^2 + 29.61*last_evaluation^3 +19.36*number_project^2
-302.6* average_montly_hours^2 +154.98* average_montly_hours^3 -81.86* time_spend_company^2
+ 31.07* time_spend_company^3 + 12.87*satisfaction_level*last_evaluation
+ 11*satisfaction_level*average_montly_hours + 20.18*satisfaction_level*time_spend_company
+ 19.62* last_evaluation*number_project + 16.74*last_evaluation*average_montly_hours
+ 13.81*last_evaluation*time_spend_company + 1.79*last_evaluation*position_sales
+ 2.34*last_evaluation*salary_low +2.78*last_evaluation*salary_medium
+ 26.31*number_project*average_montly_hours + 10.55*number_project*time_spend_company
+ 7.94*average_montly_hours*time_spend_company -0.53*time_spend_company*position_IT
-5.64*time_spend_company*position_management -2.97*time_spend_company*position_sales
-2.05*time_spend_company*salary_medium +2.47*position_management*salary_low
+ 2.29*position_management*salary_medium

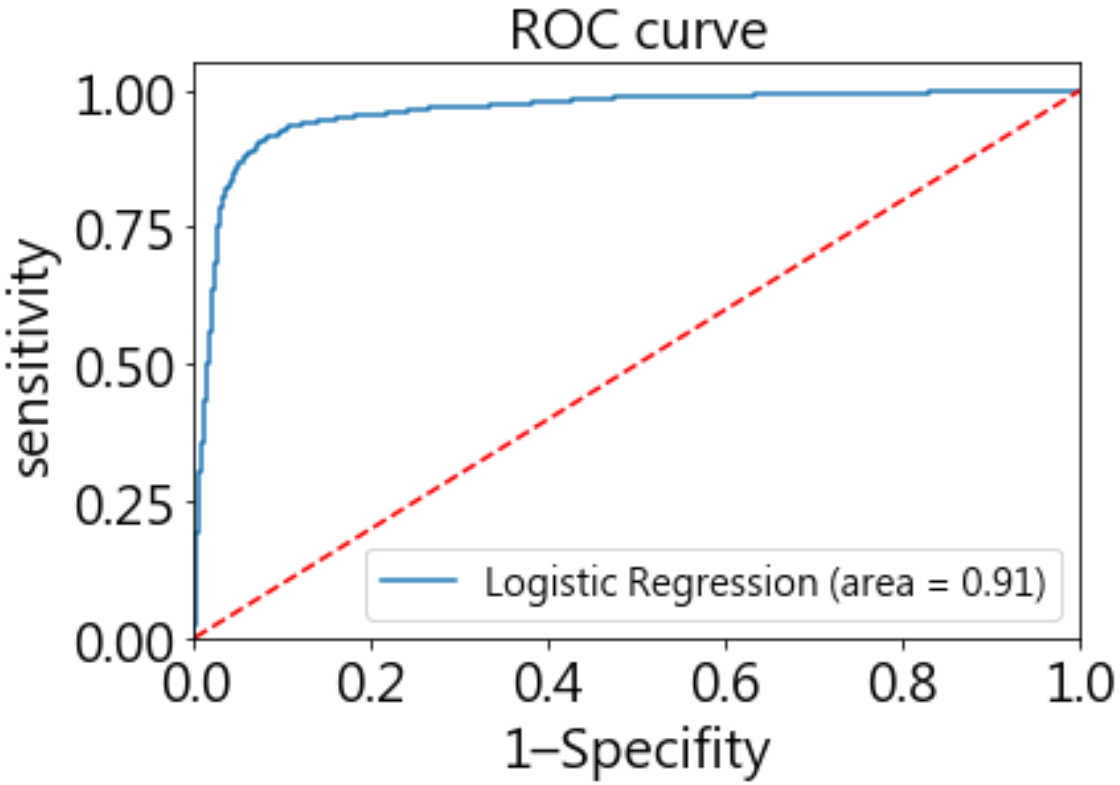
PART 4

模型評估

- 準確度
- 混淆矩陣
- ROC Curve
- 模型解釋

準確度/混淆矩陣/ROC Curve

- 訓練集配適準確度:91.03% ，測試集配適準確度:89.24%
- 敏感度=1008/(1008+64)=0.94，特異度=3009/(3009+420)=0.88
- AUC:0.91



模型解釋

模型評估

正影響(值越大，員工越容易離職)

number_project(參與過的專案數)
average_montly_hours(平均每月工時)
time_spend_company(進入公司的年數)

負影響(值越大，員工越不易離職)

satisfaction_level(員工滿意度)

- satisfaction_level (員工滿意度)又會受到last_evaluation(績效評估), average_montly_hours(平均每月工時)及time_spend_company(進入公司的年數)的影響
- last_evaluation (績效評估)主要受到number_project(參與過的專案數),average_montly_hours(平均每月工時)及time_spend_company(進入公司的年數)的影響
- 若公司想要挽留人才，可以想辦法降低員工的number_project(參與過的專案數)及average_montly_hours(平均每月工時)，以提高他們的satisfaction_level(員工滿意度) ，進而降低離職的可能性


THANK YOU FOR LISTENING