Employee Attrition Prediction

員工離職預測

國立政治大學統計碩二 許振榆

大綱

Step 01

目標

Step 02

資料敘述

- ✓ 統計敘述
- ✓ 視覺化

Step 03

建模

- ✓ 預處理
- ✓ 模型篩選

Step 04

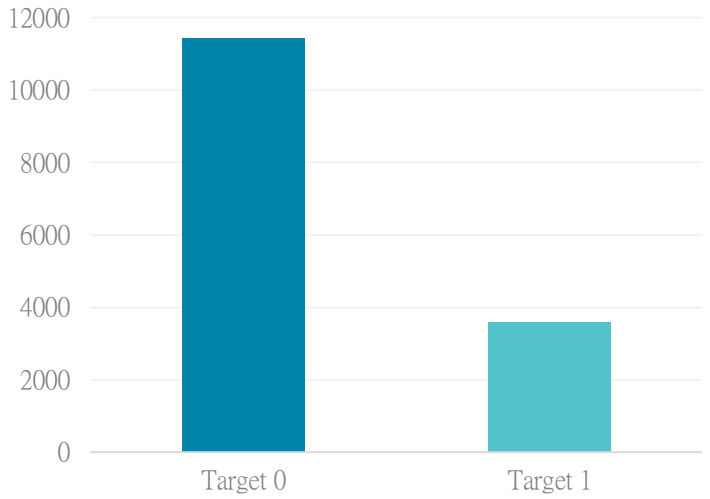
結論

- ✓ 衡量
- ✓ 建議

栽培一位員工需要時間的投入、耗費大量成本，員工作為公司的重要資產，離職將造成公司的損失，離職的問題為所有公司共通面臨的問題。

本報告透過對員工離職數據集的分析，嘗試用各種分類方法來預測有高度離職機率的員工及討論影響員工離職的關鍵因素（員工滿意度、薪資水平、平均每月工時等），根據員工離職的主要因素後，便可建議公司採取適當的措施來改善因素以挽留人才。

- 資料出處 : Kaggle
- 資料來源 : https://www.kaggle.com/ludobenistant/hr-analytics
- 資料筆數 : 14999筆員工資料(離職:未離職=23.8:76.2) 、10個變數

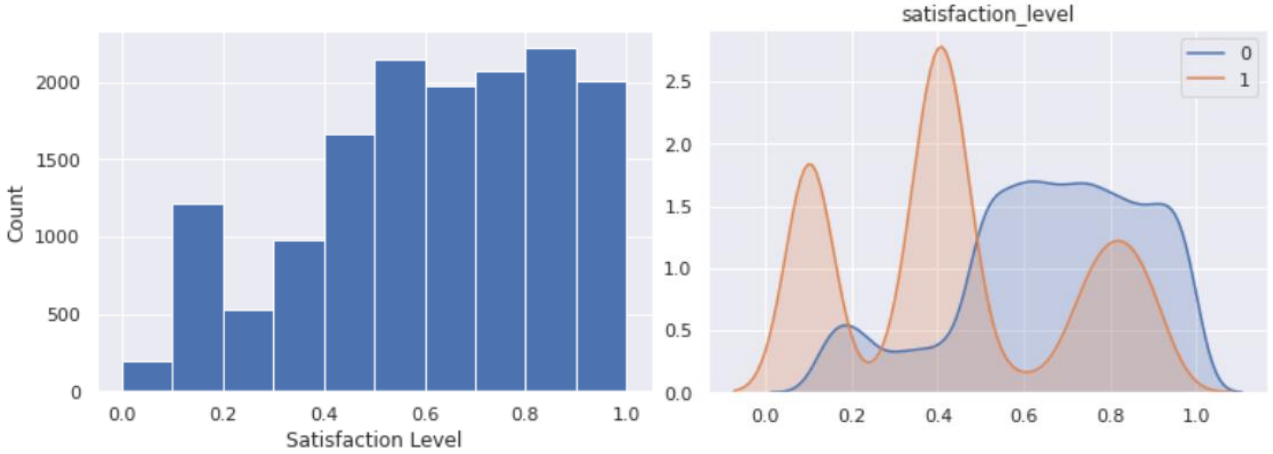


	left	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years	position	salary
0	1	0.38	0.53	2	157	3	0	0	sales	low
1	1	0.80	0.86	5	262	6	0	0	sales	medium
2	1	0.11	0.88	7	272	4	0	0	sales	medium
3	1	0.72	0.87	5	223	5	0	0	sales	low
4	1	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

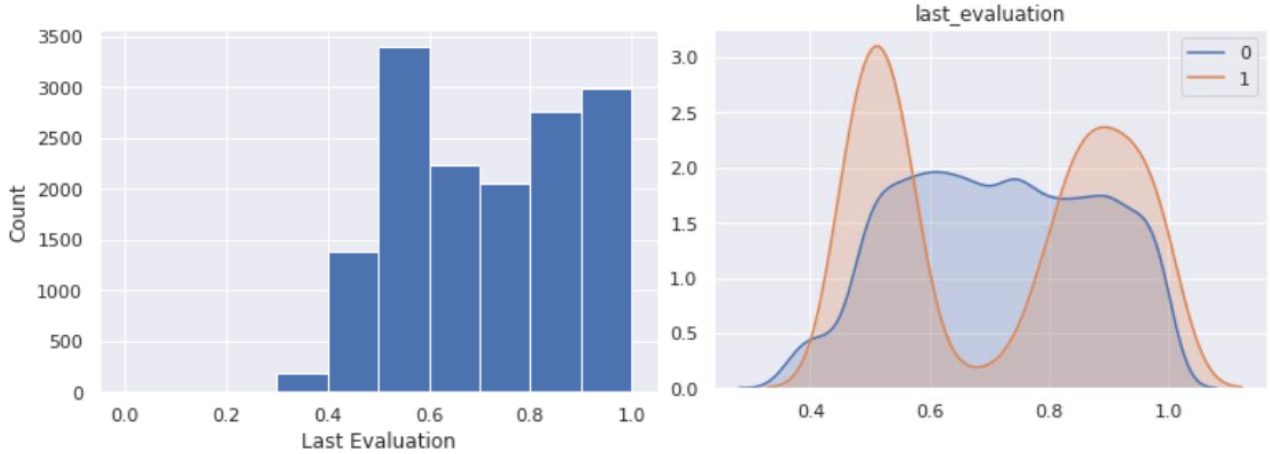
變數名稱	資料類型	變數意義
left	Boolean, 0,1	是否離職, 1代表已離職
satisfaction_level	Numeric, 0~1	員工滿意度
last_evaluation	Numeric, 0~1	績效評估
number_project	Integer, 2~7	參與過的專案數
average_montly_hours	Integer, 96~310	平均每月工時
time_spend_company	Integer, 2~10	公司年資
work_accident	Boolean, 0,1	是否有過工作意外, 1代表曾發生
promotion_last_5years	Boolean, 0,1	五年內是否升職, 1代表已升職
position	Category, 10種職位	在公司所屬部門
salary	Category, 三種水準	薪資水平

變數介紹表

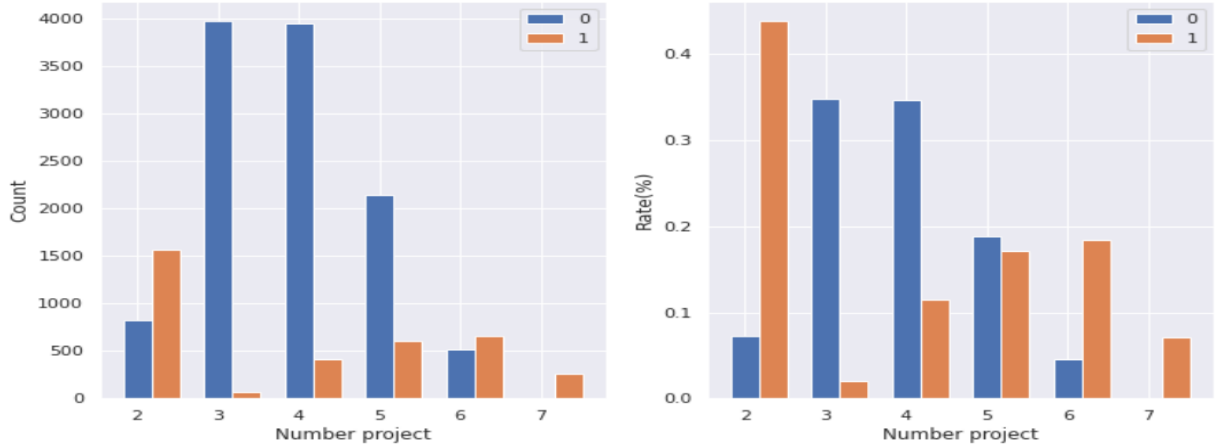
- 員工滿意度愈低 -> 離職率愈高



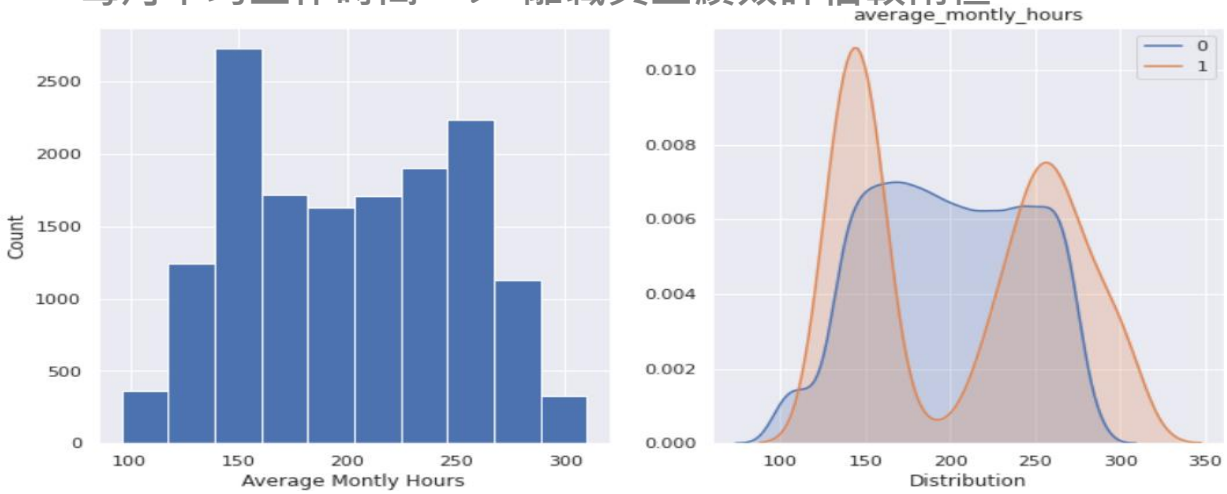
- 績效評估 -> 離職員工績效評估較兩極



- 參與過大量專案數 -> 離職率高

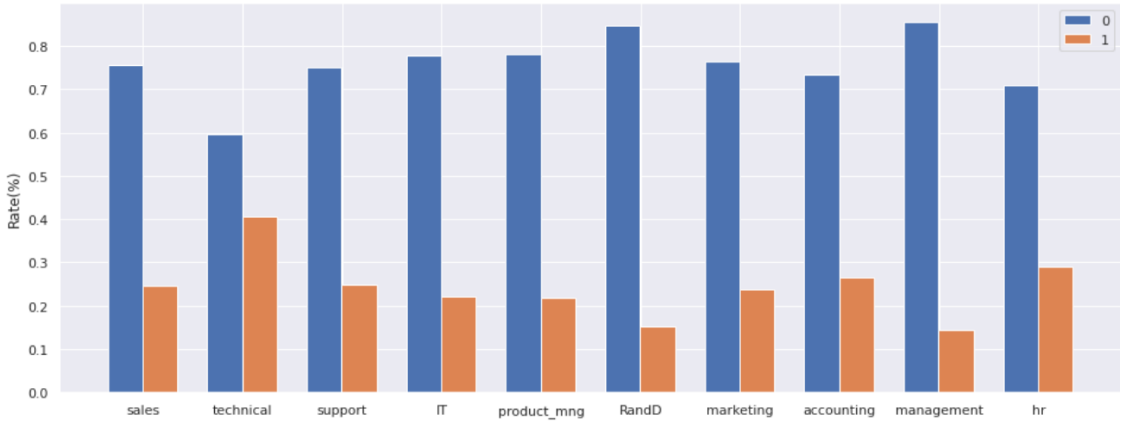


- 每月平均工作時間 -> 離職員工績效評估較兩極

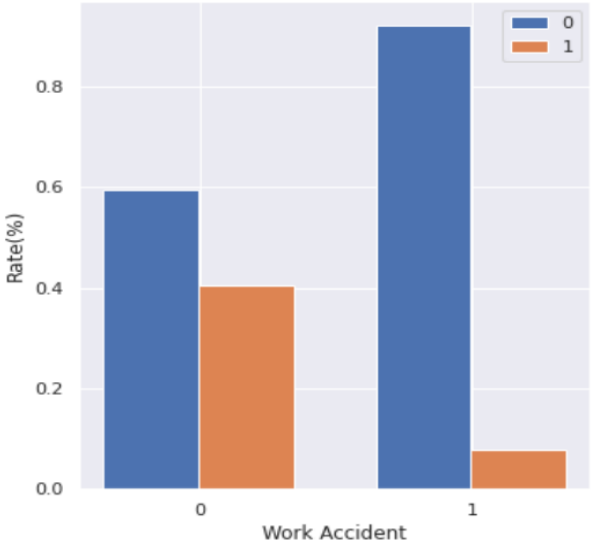


在此畫出名目型變數內各類別離職與留任的比例

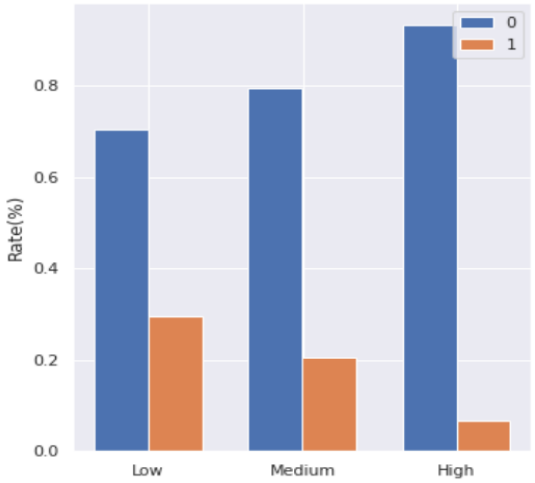
- 部門與離職並無明顯影響



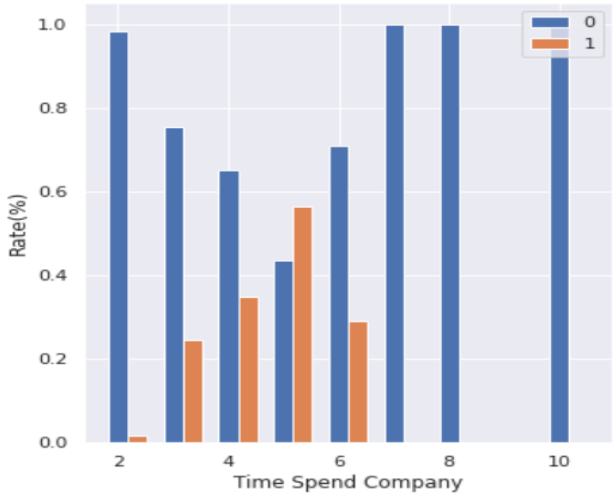
- 有無發生意外與離職並無明顯影響



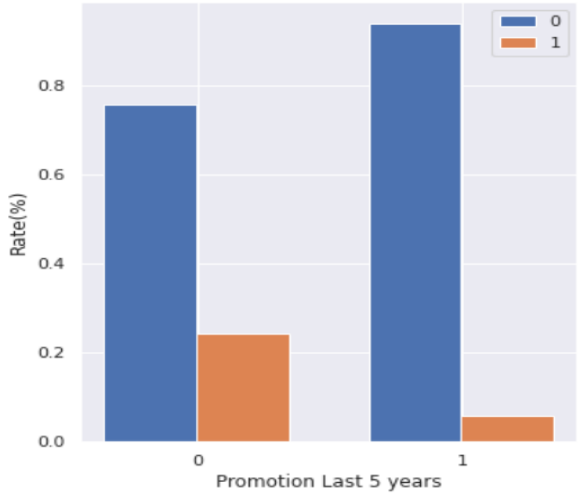
- 薪水高者離職率較低

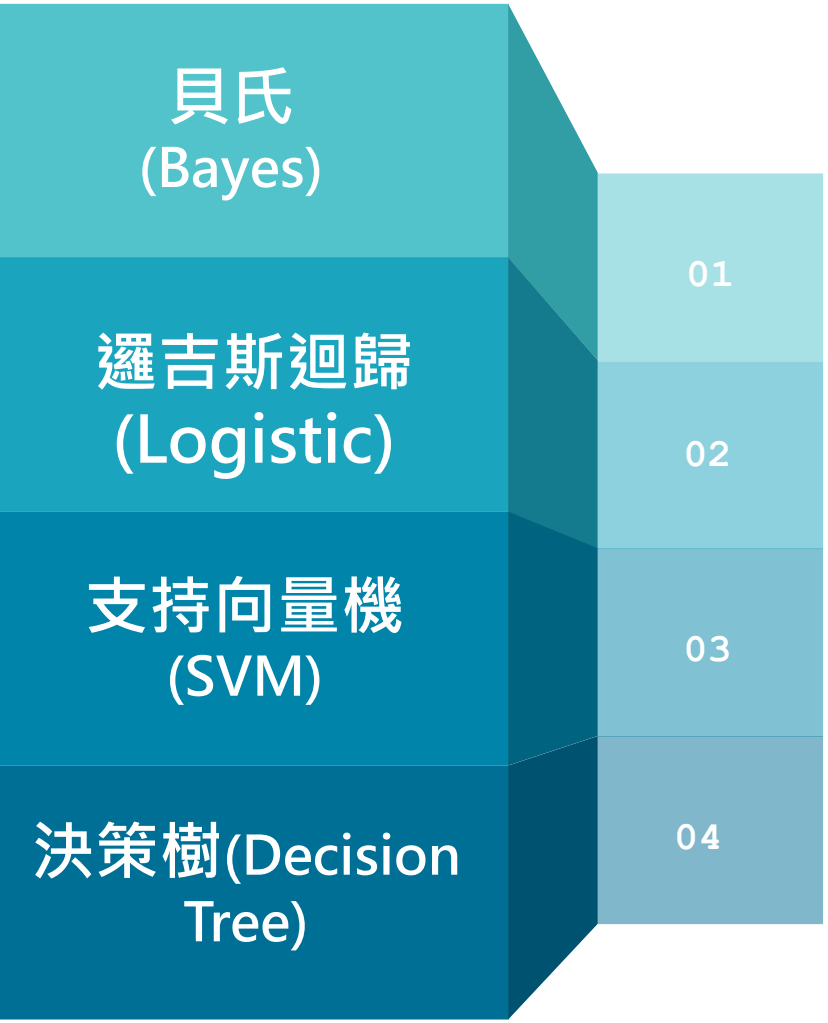


- 進入公司年數5年內離職數隨年數增加



- 5年內有無升遷與離職並無明顯影響





常見分類器

01

優)
1.建模速度快、時間短
2.適用於小資料集
缺)
1.資料滿足條件獨立

02

優)
1. 解釋性高
2. 建模速度快
缺)
1.不易處理非線性資料

03

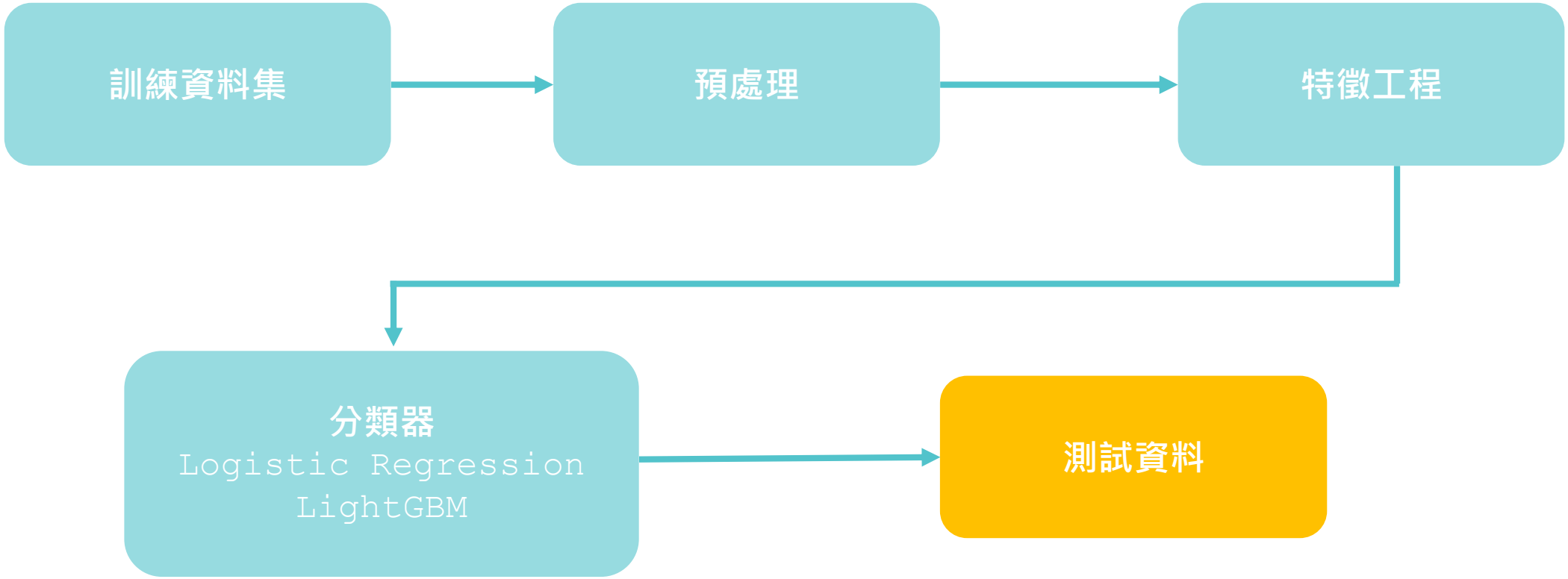
優)
1.映射到高維有較好的分類表現
缺)
1.消耗大量記憶體

04

優)
1.樹狀圖容易解釋
2.可處理類別型變數
缺)
1.容易過擬合

考量到模型解釋能力，邏吉斯迴歸能告訴我們重要特徵對反映變數的影響，首先選擇邏吉斯迴歸建立模型

建模流程



預處理

1. 數值型變數標準化

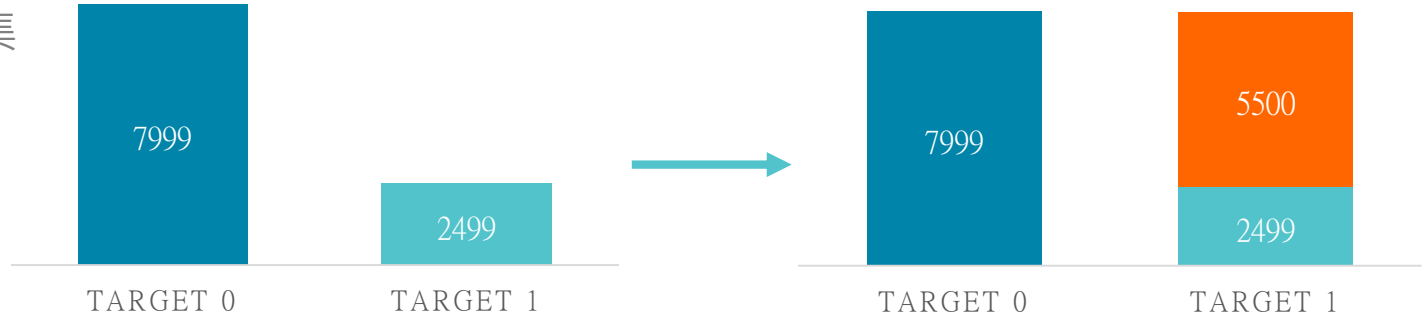
* 使數值型變數控制在平均數為0、標準差為1的相同尺度下

2. 類別型作虛擬變數

* position欄位中分為共10種，轉換9個dummy variables
* salary欄位中分為High,Medium,Low共三種，轉換為2個dummy variables
* 轉換後變數總數為18個

3. SMOTE方法模擬生成樣本

* 針對不平衡資料作處理，讓模型學習樣本數量少的類別，生成少類別使得兩類別比例相等
* 訓練資料集



特徵工程_增加變數

1. 增加交互作用項

* 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

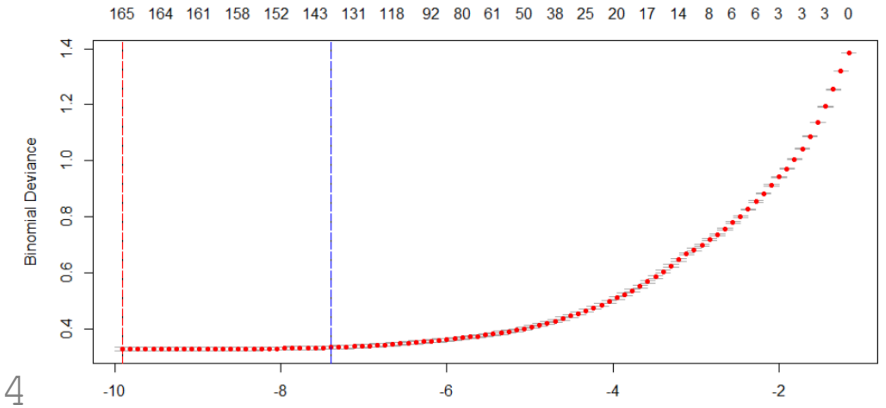
* 新增後總變數為171個

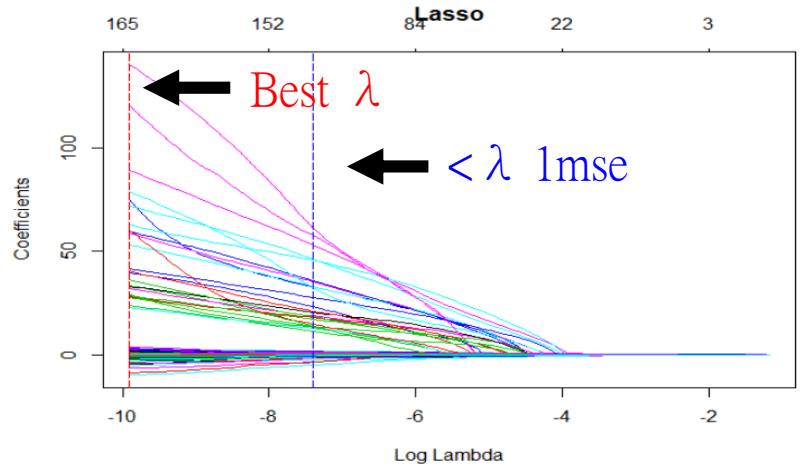
2. 增加高次項

* 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

* 新增後總變數為181個

Logistic_變數篩選





1. 使用Lasso
   - 最佳λ為0.000614
   - 變數由181個刪減為139個
   - 將剩餘139個變數以Logistic Regression建模，留下顯著P-value<0.01的變數
   - 刪除後剩餘46個變數

Lasso變數收斂情形(列出部分)

	Estimate
(Intercept)	-2.1388
satisfaction_level	-0.1208
last_evaluation	-0.2512
number_project	-0.0189
average_montly_hours	-0.3407
time_spend_company	1.9027
promotion_last_5years	.

刪除P-value>0.01的變數(列出部分)

	Estimate	Pr(>|z|)
(Intercept)	-2.363e+00	< 2e-16
satisfaction_level	-3.161e-02	0.799131
last_evaluation	-4.531e-01	0.000279
number_project	1.044e-01	0.460482
average_montly_hours	-6.493e-01	3.35e-08
time_spend_company	2.476e+00	< 2e-16
Work_accident	-1.742e+00	1.38e-14

剩餘47個變數

[1] " last_evaluation"
[2] "average_montly_hours"
[3] "time_spend_company"
[4] " Work_accident"
[5] " position_RandD "
[6] " position_product_mng "
[7] " position_sales "
[8] " position_support "
[9] " salary_high "
[10] "salary_medium "
[11] "satisfaction_level^2 "
[12] "satisfaction_level^3 “
[13] "last_evaluation^2"
[14] "last_evaluation^3"
[15] "number_project^2"
[16] "number_project^3"
[17] "average_montly_hours^2"
[18] "average_montly_hours^3"
[19] "time_spend_company^2"
[20] "time_spend_company^3"
[21] "satisfaction_level*last_evaluation"
[22] "satisfaction_level*average_montly_hours"
[23] "satisfaction_level*time_spend_company"
[24] "satisfaction_level*position_RandD"
[25] "satisfaction_level*position_technical"
[26] "last_evaluation*number_project"
[27] "last_evaluation*average_montly_hours"
[28] "last_evaluation*time_spend_company"
[29] "last_evaluation*Work_accident “
[30] "last_evaluation*position_sales"
[31] "last_evaluation*salary_high “
[32]"number_project*average_montly_hours “ …

原始變數4個
虛擬變數6個
高次變數10個
交互作用項27個

Logistic最終模型

$$\widehat{Logit(left)} =$$
$$-2.346-0.664*last\_evaluation-0.714*average\_montly\_hours +2.602*time\_spend\_company$$
$$-1.598*Work\_accident-1.222*position\_RandD-0.922*position\_product\_mng+0.246*position\_sales$$
$$+0.364*position\_support -2.752* salary\_high-0.778*satisfaction\_level^2-0.699*satisfaction\_level^3$$
$$+0.497*last\_evaluation^3 +0.919*number\_project^2-0.193*number\_project^3$$
$$+0.605* average\_montly\_hours^3 -0.498* time\_spend\_company^2 -0.647* time\_spend\_company^3$$
$$+ 0.469* satisfaction\_level*last\_evaluation+0.365 satisfaction\_level*average\_montly\_hours$$
$$+1.705*satisfaction\_level*time\_spend\_company-0.697 satisfaction\_level*position\_RandD$$
$$- 0.418*satisfaction\_level*position\_technical +0.608*last\_evaluation*number\_project$$
$$+0.456*last\_evaluation*average\_montly\_hours +0.547*last\_evaluation*time\_spend\_company$$
$$-0.126*last\_evaluation*Work\_accident+0.263*last\_evaluation*position\_sales$$
$$-1.029*last\_evaluation*salary\_high +0.496*number\_project*average\_montly\_hours$$
$$+0.451*number\_project*time\_spend\_company-0.243*number\_project*position\_marketing…$$

LightGBM

1. 簡介

* 屬於集成式學習法(Ensrmble)，建立多棵弱決策樹

2. 調校超參數

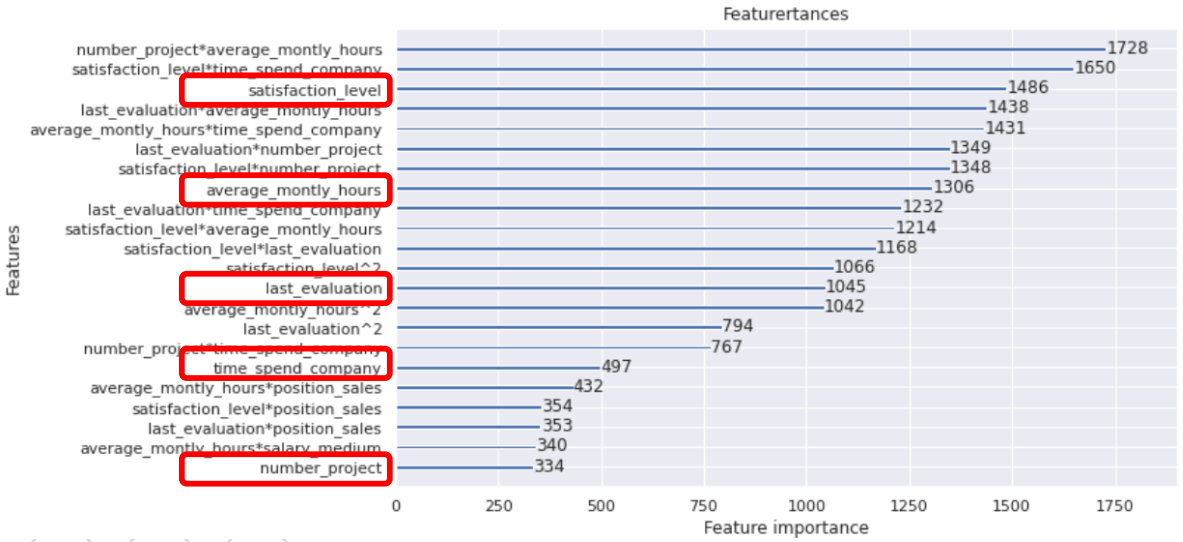
* 將訓練集做 5 摺交叉驗證(cross validation)
* num_leaves(300) -> min_data_in_leaf(20) -> num_iterations(300)
  max_bin(300) feature_fraction(1) learning_rate(0.1)

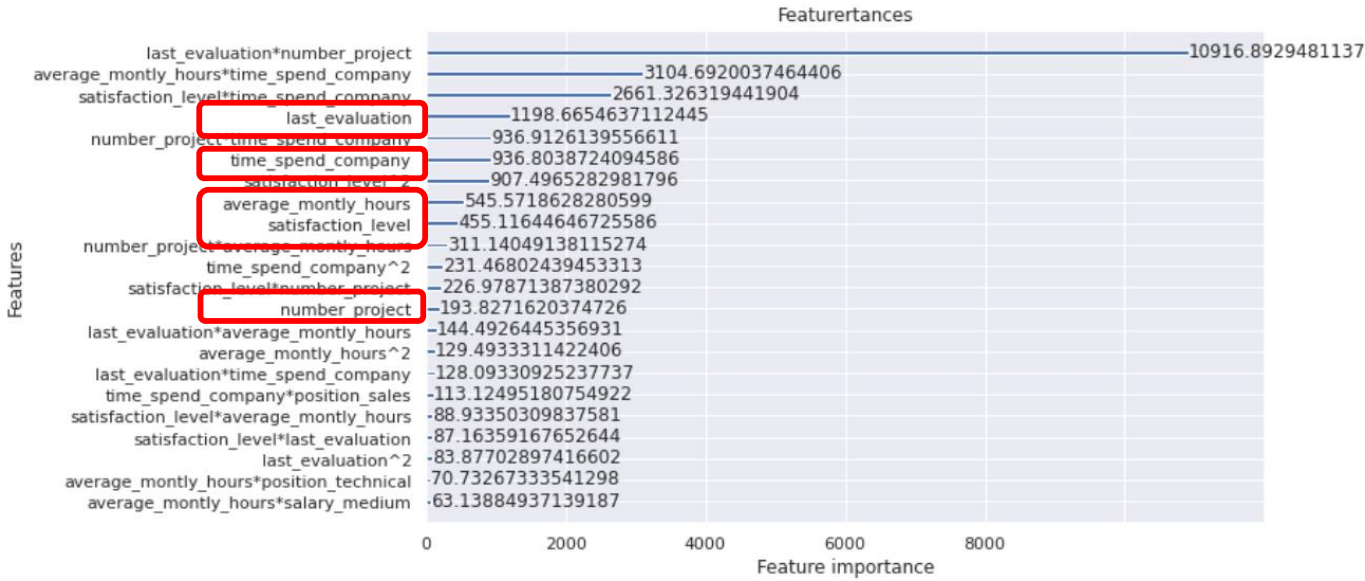
3. 訓練模型

LightGBM

4. 特徵重要性

* 左圖: 特徵使用次數
* 右圖: total gini gain
* satisfaction_level、time_spend_conpany、last_evaluation、number_project、average_monthly_hours重要特徵





分類器衡量指標

01 正確率(Accuracy) = $\frac{TP+TN}{TP+FP+FN+TN}$

不適用於不平衡資料，若全部資料預測為比例較高的類別正確率可提升，但一般稀有類別為較重要的預測目標

02 精準度(Precision) = $\frac{TP}{TP+FP}$

召回率(Recall) = $\frac{TP}{TP+FN}$

F_score = $\frac{2*precision*recall}{precision+recall}$

不平衡資料時根據類別數量多寡給定不同權重，可調整使用macro_F

03 AUC = 調整不同閥值ROC curve底下面積

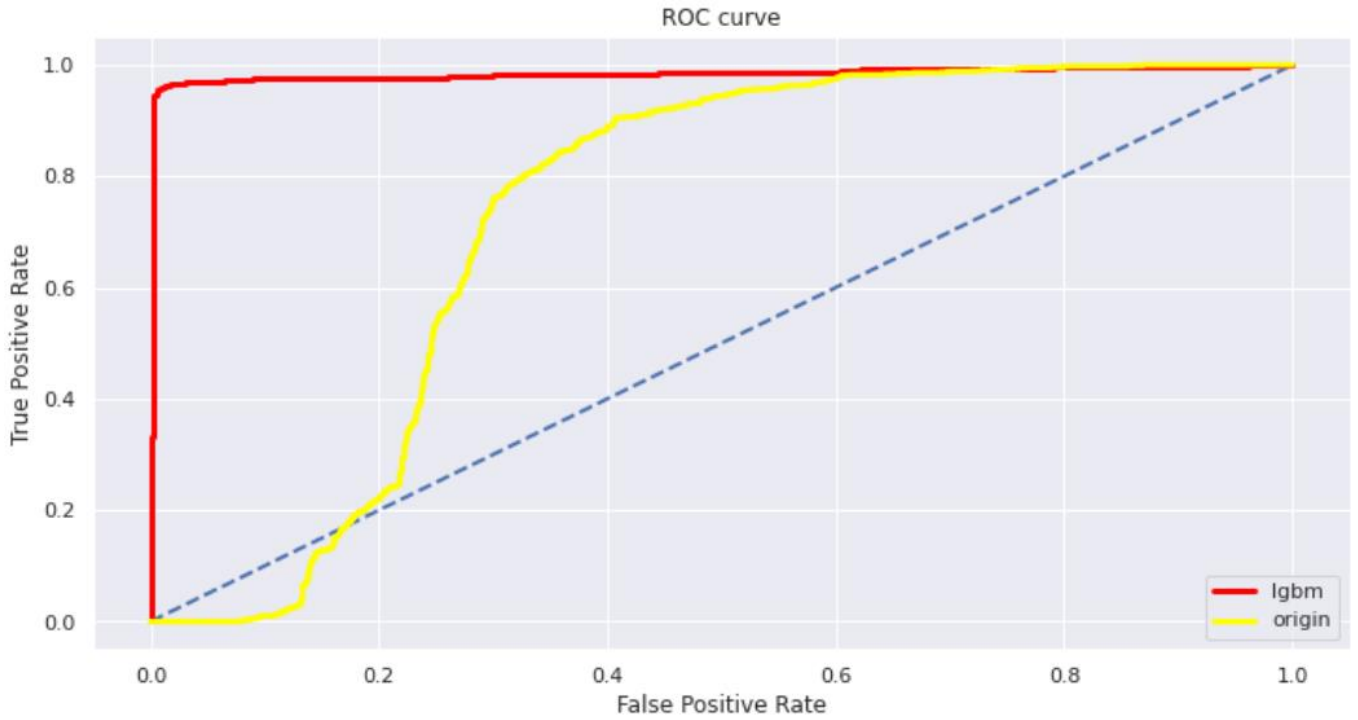
較適合用來衡量分類問題模型適用度

真實\預測	1	0
1	True Positive (TP)	False Negative (FN)
0	False Positive (FP)	True Negative (TN)

衡量測試資料集結果

- AUC比較

模型	AUC
原始	0.729
Logistic	0.981
LightGBM	0.999



經過

1. 對類別變數設虛擬變數
2. 規一化
3. 合成樣本(SMOTE)
4. 增加交互作用及高次項
5. Lasso/Stepwise特徵篩選

各模型分類結果

混淆矩陣

- Logistic

真實\預測	0	1	總共
0	3009	420	3429
1	64	1008	1072
總共	3073	1428	4501

- LightGBM

真實\預測	0	1	總共
0	1134	8	1142
1	0	358	358
總共	1134	366	1500

各類別分類指標

precision	0.706
recall	0.941
Accuracy	91.03%
F1-score	0.807

precision	0.978
recall	1
Accuracy	99.47%
F1-score	0.989

促進離職特徵
(值愈大，愈容易離職)

time_spend_company (公司任職時間)

影響留職特徵
(值愈小，愈容易離職)

average_montly_hours (平均每月工時)
number_project (參與專案數)
satisfaction_level (員工滿意度)
last_evaluation(績效)

建議

- ✓ 若公司想要挽留人才，衡量員工的average_montly_hours(平均每月工時)，給予能力匹配的工作量
- ✓ 年資第四、五年為員工離職高峰期，優秀員工給予具吸引力的福利
- ✓ number_project(參與專案)、satisfaction_level(員工滿意度)、 last_evaluation(績效)低，反映員工離職傾向，探究可能原因為公司內部因素或個人原因

Thanks