110305075 企管三 繆孟珊
109305027 企管四 顏　絜
108702021 心理四 何瑀芹
112752017 心碩一 詹佑茗
113752005 心碩一 張雅雯

Ch7
員工離職傾向分析

個案背景

- 面臨問題：近年公司招募的應屆畢業生離職率偏高（入職三年內流失50%）
- 離職原因：
  - 認知偏差
  - 適應性差
  - 追求高薪資福利
  - 內部管理原因
- 解決方案：建立模型在招聘時預測新鮮人入職後離職率，以作為招聘參考依據
- 目標：改善離職現象，減低招聘成本
- 數據分析方法：機器學習（Boosting、Random Forest）

所需數據

- **個人基本資料**
  - 姓名、性別、工作單位、入職年份、學歷、畢業院校、專業、職稱、職業資格、是否黨員、錄用時崗位級別、是否有晉升
- **招聘時綜合評鑑分數**
  - 勝任力測評
    - 抗壓能力、外向性、社交能力、心理感受性、創新、敬業、情緒穩定性、嚴謹性、完美主義傾向
  - 基本素質測評
    - 言語理解、數學、邏輯、常識、成就導向
- **入職後一年內離職情形**(是、否)

資料範例

- 個人基本資料、入職後一年內離職情形

序號	姓名	性別	工作單位	工作單位類別	入職年份	學歷	畢業院校	專業	職稱	職業資格	是否黨員	錄用時崗位級別	是否有晉升	是否1年內離職
1	梁**	男	N分公司	G類	2009	碩士	華南師範大學	人力資源管理	無	無	是	10	1	0
2	李**	女	N分公司	G類	2009	碩士	華南師範大學	應用心理學	無	無	否	10	1	0
3	傅**	男	N分公司	G類	2009	本科	廣東技術師範學院	通信工程	助理工程師	無	是	12	1	0
4	葉**	女	N分公司	G類	2010	碩士	廣東商學院	企業管理	無	無	是	10	0	0

資料範例(續)

- 招聘時綜合評鑑分數

  - 基本素質測評

序號	言語理解	數學	邏輯	常識	成就導向
1	8	15.5	22.6	3.5	7.9614
2	9	11.5	17.8	4.2	6.01003
3	10	9.5	12	4.9	5.61976
4	9	12.5	15.9	4.9	7.57113

  - 勝任力測評

序號	抗壓能力	靈活性	影響性	支配性	外向性	社交能力	心理感受性	創新	敬業	情緒穩定性	嚴謹性	完美主義傾向
1	4.46076	5.17845	6.86959	7.8381	6.28064	6.48218	4.57265	6.19874	5.36425	4.07274	3.05942	6.54208
2	4.46076	3.54126	4.20825	5.63132	5.13687	5.85147	6.0989	4.88433	7.65804	6.29585	5.16646	4.11819
3	6.86273	5.72419	6.86959	6.36691	6.6619	5.22075	5.33578	6.63687	4.9055	5.85123	5.16646	4.11819
4	4.46076	6.26992	7.53493	8.57369	6.28064	5.22075	6.0989	5.7606	5.36425	5.85123	5.16646	4.60297

機器學習介紹

- 相對傳統統計方法優點:
  - 更高預測精度
  - 更高自變量容許度
  - 無需自行設定自變量重要性、篩選自變量
  - 可計算自變量重要性
  - 可迭代加強
- 採用算法:
  - Boosting
  - Random Forest
  - 皆為基於決策樹(弱分類器)的強分類器，可二元判定因變量結果
  - 兩種算法可以相互應證、比較，實際運用時選取準確率較高算法

算法介紹

Boosting

- 分類器的一種
- 決策樹的加強版本(強分類器)
- 具有自適應特點
- R默認可生成50棵決策樹
- 優點:
  - 預測準確率高
  - 無過度擬和問題
  - 不挑惕自變量類型、數量
- 缺點:
  - **易受奇異點或離群值影響**

Random Forest

- 分類器的一種
- 決策樹的加強版本(強分類器)
- 決策樹每個節點的變量在隨機選出的少數變量中產生，每棵決策樹依據的數據與節點產生皆為隨機，串連所有隨機生成的決策樹形成森林
- R默認可生成500棵決策樹
- 優點:
  - 預測準確率高
  - 無過度擬和問題
  - 不挑惕自變量類型、數量
  - **適合使用於自變量多的大數據**

分析方法

- 因變量設定：預測新鮮人離職率→「是否一年內離職」
- 自變量：其餘因子
- 分析步驟：
  - 建立模型
  - 檢驗預測效果
  - 檢視模型對於「各自變量重要性」分析
  - 應用：將新鮮人資料帶入模型，預測入職後離職率，作為甄選參考之一

分析步驟:建立模型

- Random Forest:因子類變量不得超過53個類別，須將原始數據進行轉換
  - 畢業院校 → 依學校所處省份歸類
  - 專業 → 依學科分類

序號	畢業院校	畢業院校地區	專業	專業類別
1	華南師範大學	廣東	人力資源管理	管理學
2	華南師範大學	廣東	應用心理學	理學
3	廣東技術師範學院	廣東	通信工程	工學
4	廣東商學院	廣東	企業管理	管理學

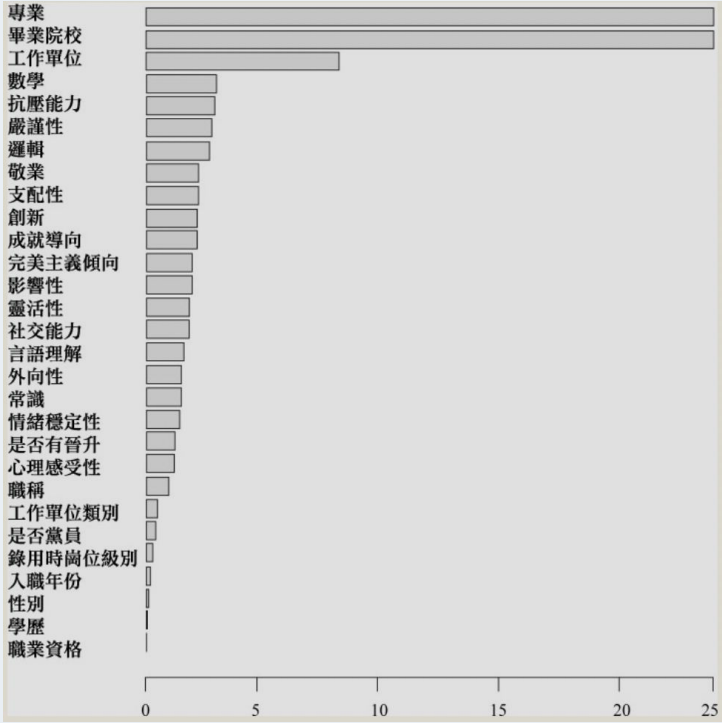
分析步驟:檢驗預測效果

- 依據原始數據預測
- 兩種算法誤判率皆為0

實際 \ 預測	離職	在職
在職	0	1 217
離職	242	0

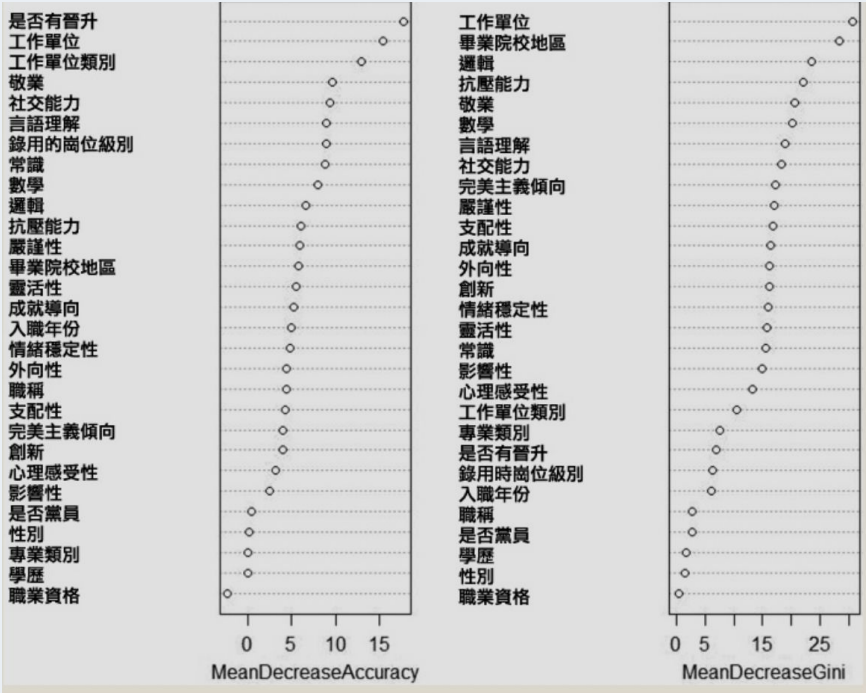
分析步驟:「各自變量重要性」分析

- Boosting
- 重要因素:
  - 專業
  - 畢業院校
  - 工作單位



分析步驟:「各自變量重要性」分析

- Random Forest
- 重要因素:
  - MeanDecreaseAccuracy:
    將一個變量取值變為隨機數，使模型預測準確性降低的程度
    - 是否晉升
    - 工作單位
    - 工作單位類別
  - MeanDecreaseGini:
    變量對決策樹每個節點觀測值異質性的影響
    - 工作單位
    - 畢業院校地區
    - 邏輯能力



資料問題

- 跑敘述統計
- 確認無空缺、誤植之資料。

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
姓名	性别	工作单位	工作单位类别	入职年份	学历	毕业院校地区	专业类别	职称	职业资格	是否党员	言语理解	数学	逻辑	常识	成就导向	抗压能力	灵活性	影响性	支配性	外向性	社交能
梁**	男	N分公司	G类	2009	硕士	广东	管理学	无	无	是	8	15.5	22.6	3.5	7.9614	4.46076	5.17845	6.86959	7.8381	6.28064	6.48
李**	女	N分公司	G类	2009	硕士	广东	理学	无	无	否	9	11.5	17.8	4.2	6.01003	4.46076	3.54126	4.20825	5.63132	5.13687	5.85
傅**	男	N分公司	G类	2009	本科	广东	工学	助理工程师	无	是	10	9.5	12	4.9	5.61976	6.86273	5.72419	6.86959	6.36691	6.6619	5.22
叶**	女	N分公司	G类	2010	硕士	广东	管理学	无	无	是	9	12.5	15.9	4.9	7.57113	4.46076	6.26992	7.53493	8.57369	6.28064	5.22
韩**	女	N分公司	G类	2011	博士	广东	理学	无	无	是	15	24.8	25.6	10.8	4.83921	5.06125	2.44979	4.20825	4.89573	5.51813	4.90
骆**	男	N分公司	G类	2011	硕士	广东	管理学	无	无	是	15	20.5	21.9	9.6	5.22948	6.26223	5.17845	4.87359	6.73471	5.51813	5.22
姚**	男	N分公司	G类	2011	硕士	广东	工学	无	无	是	10.5	22.5	24.5	10.8	5.37583	7.11293	4.63272	4.87359	4.89573	4.75561	5.53
余**	男	N分公司	G类	2011	硕士	云南	文学	无	无	是	13.5	22.5	17.9	10.8	7.57113	5.96199	4.08699	5.53892	7.8381	5.13687	5.85
姚**	男	N分公司	G类	2011	本科	广东	理学	无	无	否	15	22.5	16.5	9.6	5.22948	5.3615	7.36138	6.20426	5.26353	5.89939	5.85
蔡**	男	N分公司	G类	2011	本科	广东	工学	无	无	否	19.5	28.5	29	12	7.18085	5.3615	7.36138	8.20026	7.4703	6.28064	5.51
苏**	女	N分公司	G类	2011	本科	广东	文学	无	无	否	18	24.5	37	12	3.6196	3.56002	8.45285	2.87758	2.68896	3.61184	3.95
林**	女	N分公司	G类	2011	本科	吉林	文学	无	无	否	13.5	22.8	24.2	9.6	3.66838	5.81187	2.44979	4.87359	6.36691	3.9931	5.22
黄**	女	D分公司	B类	2011	本科	境外	管理学	无	无	否	10.5	20.8	17.1	8.4	2.88784	5.3615	2.99553	5.53892	5.26353	7.42441	7.1

敘述統計

Descriptive Statistics

	Valid	Missing	Mean	Std. Deviation	Minimum	Maximum
序号	1459	0	730.000	421.321	1.000	1459.000
姓名	1459	0				
性别	1459	0				
工作单位	1459	0				
工作单位类别	1459	0				
入职年份	1459	0	2010.020	0.795	2009.000	2012.000
学历	1459	0				
毕业院校地区	1459	0				
专业类别	1459	0				
职称	1459	0				
职业资格	1459	0				
是否党员	1459	0				
言语理解	1459	0	9.658	3.437	1.000	22.500
数学	1459	0	13.247	5.630	1.000	28.500
逻辑	1459	0	20.736	5.350	1.500	37.000
常识	1459	0	6.156	2.654	0.700	12.600
成就导向	1459	0	4.951	1.584	0.546	10.000
抗压能力	1459	0	5.211	1.598	0.257	10.000
灵活性	1459	0	4.976	1.636	0.267	10.000
影响性	1459	0	5.036	1.611	0.216	10.000
支配性	1459	0	5.658	1.710	0.482	10.000
外向性	1459	0	5.061	1.463	0.181	8.568
社交能力	1459	0	4.988	1.536	0.175	9.005
心理感受性	1459	0	4.729	1.485	0.757	9.151
创新	1459	0	4.879	1.545	0.065	9.704
敬业	1459	0	5.067	1.553	0.318	9.952
情绪稳定性	1459	0	4.944	1.531	0.071	8.964
严谨性	1459	0	4.562	1.450	0.110	10.000
完美主义倾向	1459	0	5.042	1.611	0.240	10.000
录用时岗位级别	1459	0	11.813	1.282	9.000	16.000
是否有晋升	1459	0	0.249	0.432	0.000	1.000
是否1年内离职	1459	0	0.166	0.372	0.000	1.000

Note. Not all values are available for *Nominal Text* variables

Boosting分析結果

Boosting Classification ▼

Boosting Classification ▼

Trees	Shrinkage	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy
12	0.100	934	234	291	0.833	0.856

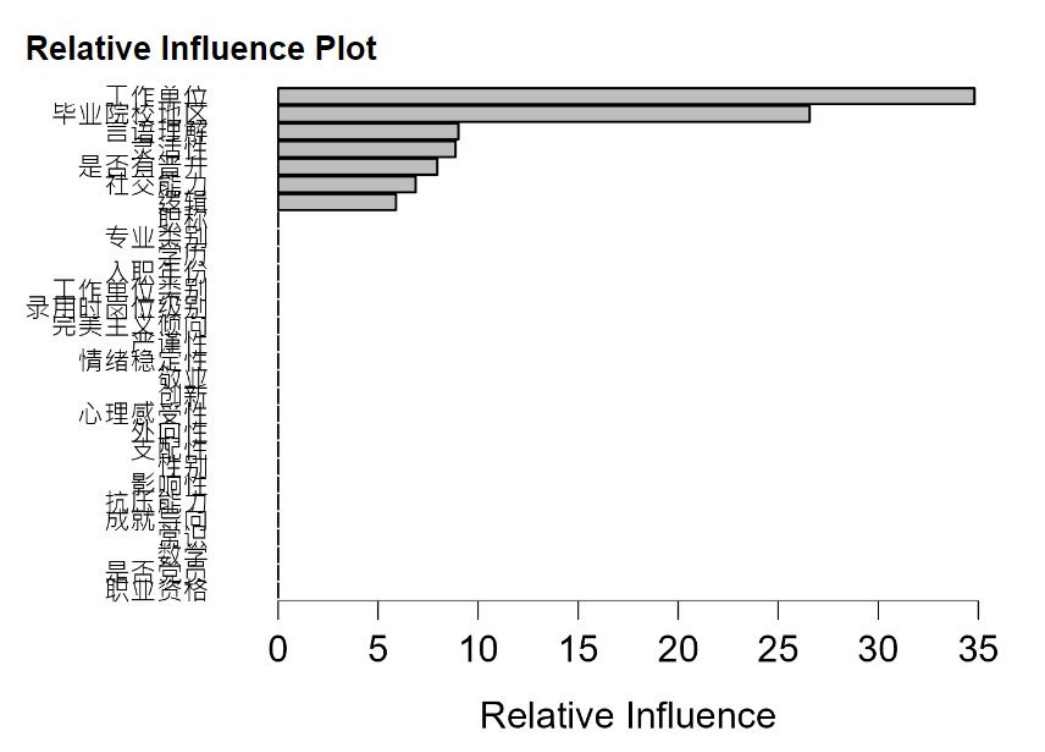
Note. The model is optimized with respect to the *out-of-bag accuracy*.

Data Split

Confusion Matrix

		Predicted
		0
Observed	0	249
	1	42

Boosting分析結果



1. 工作單位
2. 畢業院校地區
3. 言語理解
4. 靈活性
5. 是否有晉升

Random forest分析結果

Random Forest Classification

Random Forest Classification

Trees	Features per split	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy	OOB Accuracy
202	5	934	234	291	0.82906	0.81787	0.00000

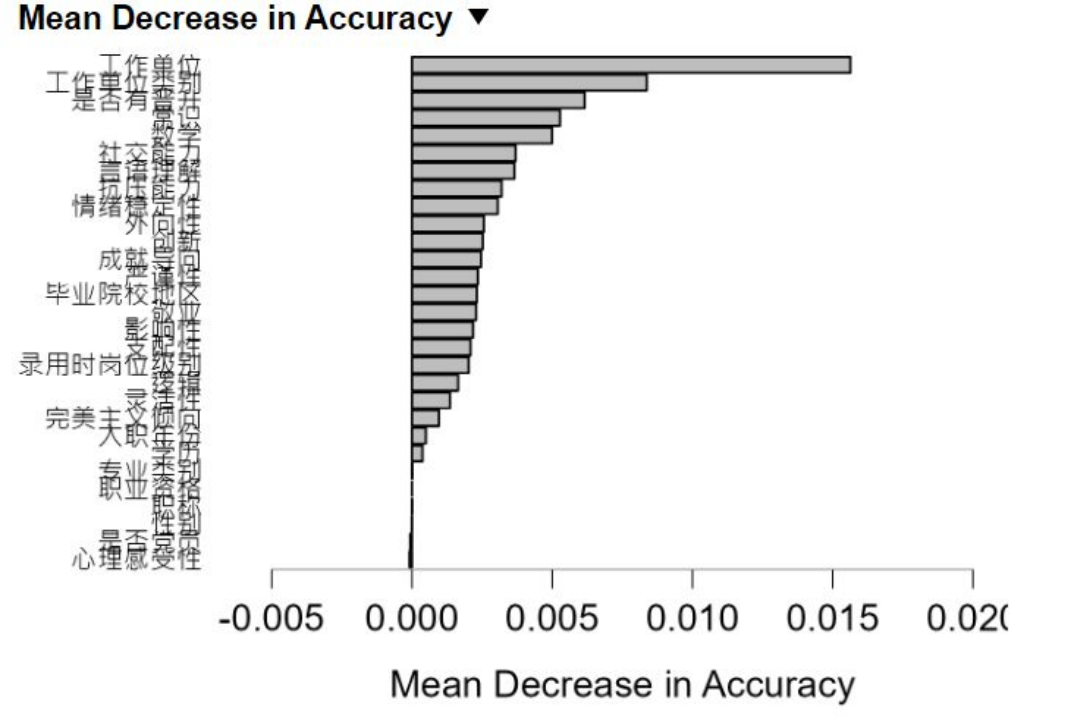
Note. The model is optimized with respect to the *out-of-bag accuracy*.

Data Split

Confusion Matrix

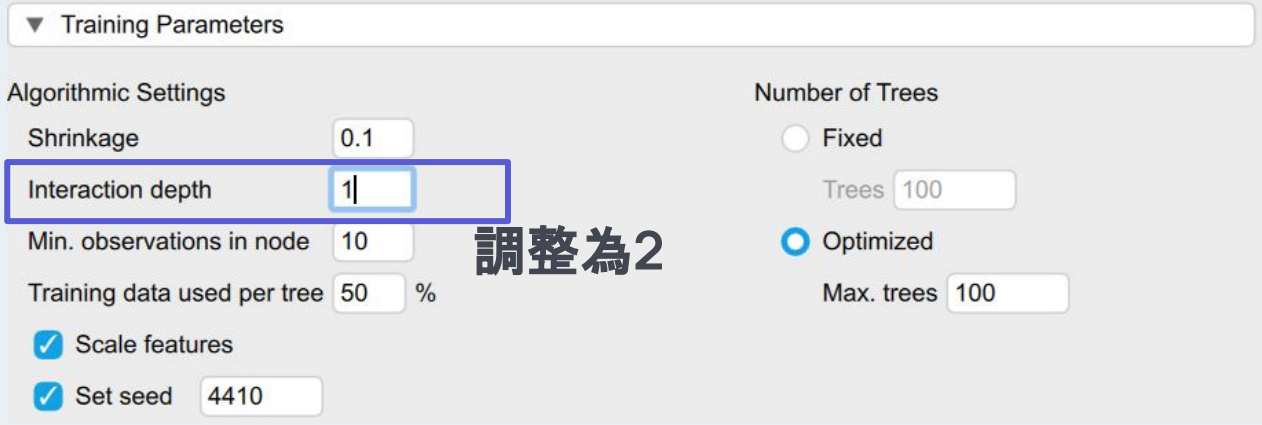
		Predicted	
		0	1
Observed	0	238	0
	1	53	0

Random forest分析結果



1. 工作單位
2. 工作單位類別
3. 是否有晉升
4. 常識
5. 數學

- **JASP跑boosting的confusion matrix缺少predicted 1:**
  - 推測1:資料不平衡(0太多1太少)，使模型傾向預測較多樣本的類別
  - 推測2:模型設置不佳＞＞調整樹的深度



調整深度後boosting結果

Boosting Classification

Trees	Shrinkage	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy
12	0.10000	934	234	291	0.83761	0.85223

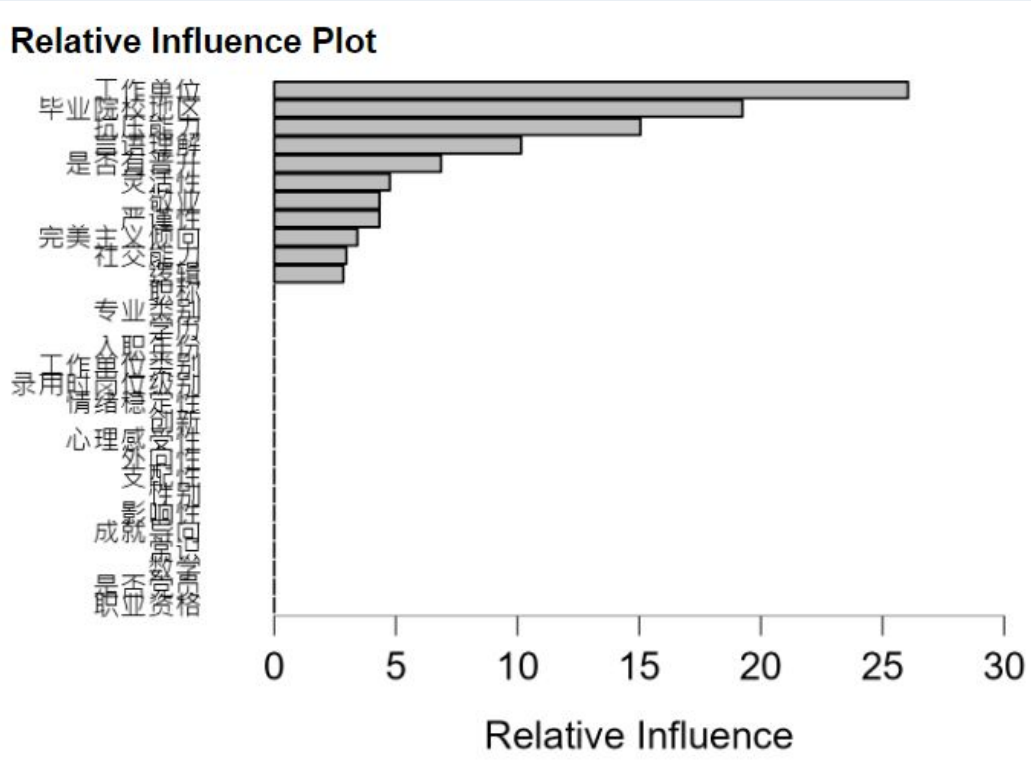
Note. The model is optimized with respect to the *out-of-bag accuracy*.

Data Split

Confusion Matrix

		Predicted	
		0	1
Observed	0	248	1
	1	42	0

調整深度後boosting結果



1. 工作單位
2. 畢業院校地區
3. 抗壓能力
4. 言語理解
5. 是否有晉升

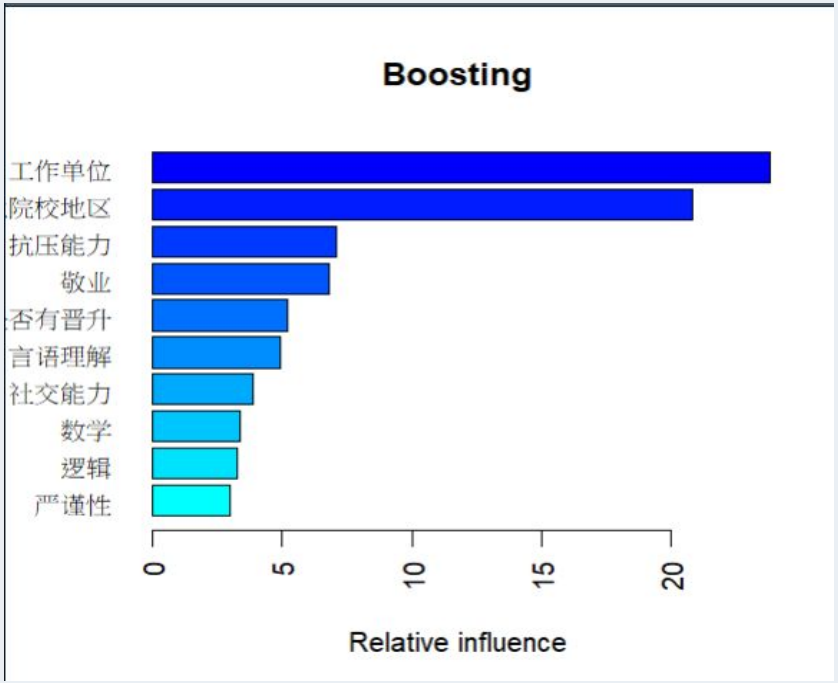
使用R的結果：混淆矩陣與課本相同

Boosting

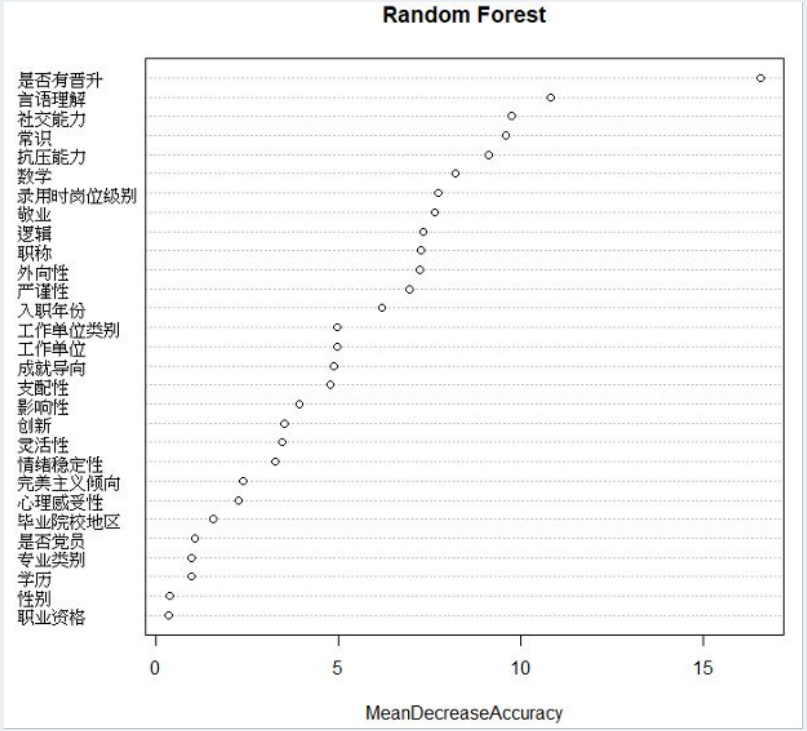
Random forest

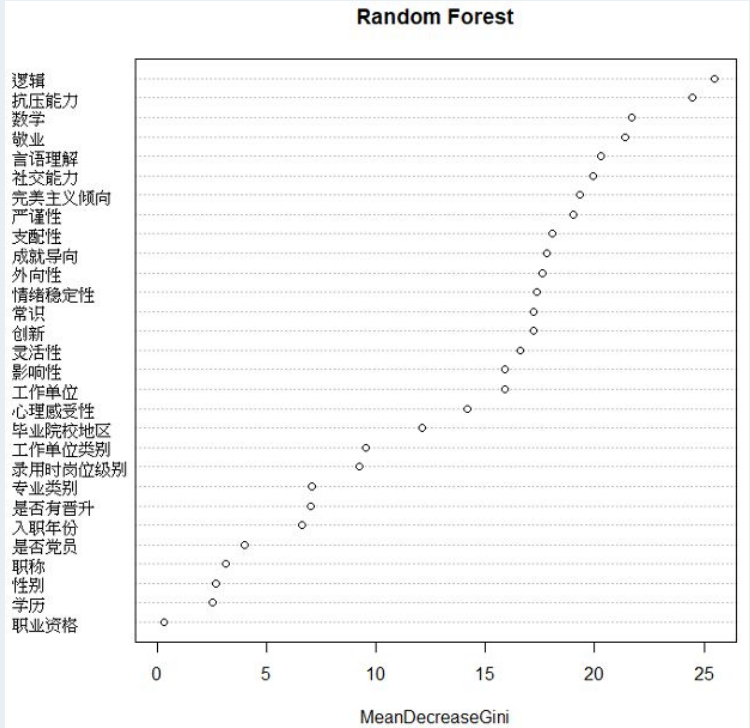
```
> table(dta$是否1年内离职, bp$class)

        在职 离职
  在职 1217    0
  离职    0  242
> table(dta2$是否1年内离职, rp)
      rp
        在职 离职
  在职 1217    0
  离职    0  242
>
```

Boosting (using R)



Random forest (using R)





Boosting vs Random Forest 分析結果比較

Relative influence	Boosting in JASP (已調整樹深度2)	Random Forest (MeanDecrease Accuracy) in JASP	Boosting in R	Random Forest (MeanDecreaseAccuracy) in R	Random Forest (MeanDecrease Gini) in R
1	**工作單位**	**工作單位**	**工作單位**	**是否有晉升**	邏輯
2	畢業院校地區	工作單位類別	院校地區	**言語理解**	**抗壓能力**
3	**抗壓能力**	**是否有晉升**	**抗壓能力**	社交能力	**數學**
4	**言語理解**	**常識**	**敬業**	**常識**	**敬業**
5	**是否有晉升**	**數學**	**是否有晉升**	**抗壓能力**	**言語理解**

建議實務作法

- 工作單位
  了解部門性質(e.g.業界相似部門的平均流動率)、部門氛圍、主管領導風格，與該部門主管溝通回報狀況，再判斷HR部門如何協助
- 抗壓能力、言語理解
  - 甄選優化針對這兩個項目的測驗鑑別度，並提升在綜合評量標準中的權重
  - 試用期著重觀察新聘人員在這兩個方面的表現
- 是否有晉升
  - 檢視原始資料中的晉升者特性，評估是否有績效考核與升遷公平性問題
  - 透過問卷或面談調查員工對於升遷公平性的認知，以便後續制度改善

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution

Thanks!