大數據要我別聘你

110305075 企管三 繆孟珊 109305027 企管四 顏 絮

108702021 心理四 何瑀芹 112752017 心碩一 詹佑茗

113752005 心碩一 張雅雯

檢查與整理資料

調整變項	調整內容
AgeRange	● Year1:將所有年齡變項合併為一欄，並將 AgeRange_<=20coding為1、AgeRange_21-30 coding為2、AgeRange_31-40coding為3，並將整併後資料為 0的資料點coding為4（假設年齡區間為 40歲以上） ● Year2: 同第一年作法
AreaName	● 將所有AreaName資料合併為一欄，並將 AreaName_EastChina coding為1、AreaName_NorthChina coding為2、AreaName_NortheastChina coding為3、AreaName_NorthwestChina coding為4、AreaName_SouthCentralChina coding為5、AreaName_SouthwestChina coding為6
OnBoardMonth	● 將所有月份的OnBoardMonth合併為一欄，並依照月份 coding為1-12
OriRecruitCat	● 將所有OriRecruitCat欄位合併，並將原始欄位為 OriRecruitCat_B coding為2、OriRecruitCat_C coding為3

檢查與整理資料

調整變項	調整內容
OriContractCat	● 在Year1中發現資料有很多為0且兩年之間差異太大不合理，因此不分析此變項
DormAreaMost	● 將所有DormAreaMost資料合併為一欄，並依照原始欄位A -O coding為1-15
ShiftCateoryMost	● 將所有ShiftCateoryMost合併為一欄，並依照原始欄位A -G coding為1-7

員工留任(retention)的操作型定義

- 留任定義為未離職之員工，也就是Quit變項=0

如何篩選出15個分析指標

- 由於ID只適合用在不同資料庫合併時作為 primary key，因此，在完成資料庫合併後，可以在分析前直接將其捨去。
- Quit為效標變項，因此不適合放在自變量中。同時我們發現 totalday是離職的人才有數值，若還在職則為空值，因此不適合放入比較中。
- OnBoardPeakSeason 與 OnBoardMonth_01-12 相比之下，OnboardMonth_01-12 包含 OnboardPeakSeason，且資料本身更為仔細，因此選擇留下 OnboardMonth_01-12。
- LeavePeakSeason變量可以了解人才大量外流的真實情況是否真的是因為 產業高峰期而導致，以利後續針對性的解決方案。
- 在考慮員工住宿環境變量時，我們認為個人所在宿舍區的環境狀況比住過多少宿舍更為重要。因此，我們認為DormAreaMost_A-O比DormAreaCount更具客觀性，也能提高模型的準確性。

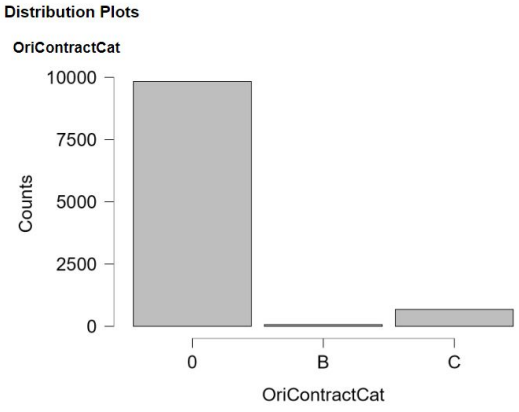
如何篩選出15個分析指標

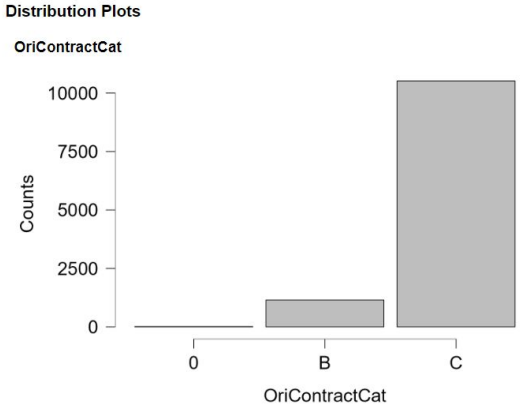
- 認為住宿比例不會影響離職留任意願，因此刪除 LiveDormRatio。
- 考慮到工作中提高獎金的誘因，我 们認為薪資和獎金的多寡及變化相對於原始薪資和薪資變化更為重 要 。因 此 ， 我 們 在 SalaryBonusMean 和 SalaryMean 中 選 擇 了 SalaryBonusMean；在 LastMonthSalaryBonus和LastMonthSalary中選擇了LastMonthSalaryBonus。
- 選擇保留LastMonthSalaryPct和LastMonthSalaryBonusPct，因為我們認為在短時間 內離職，很可能是由於短期內的變化迅速導致的。
- 愈 了 解「較 多 加 班 機 會」理 由 是 否 存 在 且 重 要，選 擇 保 留 WorkOnRestDay、ShiftRation、WorkHourMean。
- 將AgeRange合併成一個變項；將 AreaName合併成一個變項，用以了解工廠年齡分布及工廠員工的來源。

如何篩選出15個分析指標

- 在OriRecruitCat_A-D及OriContractCat_A-C中，我們選擇移除OriContractCat_A-C是因為兩年的資料差距過大，且在Year1中含有大量的0。
- ShiftCategoryMost_A-G也納入模型，因為許多資料都會在部門、生 產線上有所差異，因此先將其納入模型。

➢ 總體而言，我們選擇的變數包含：Gender, LeavePeakSeason, Turn, SalaryBonusMean, LastMonthSalaryBonus, LastMonthSalaryBonusPct, WorkOnRestDay, ShiftRation, WorkHourMean, AgeRange, AreaName, OnBoardMonth_01-12, OriRecruitCat_A-D, DormAreaMost_A-O, ShiftCategoryMost_A-G.





regular specification（enter進入方式）

由「標準化係數」（standardized）看個別因素重要性排序

重要性排序	Year 1		Year 2	
	因素	標準化係數	因素	標準化係數
1	離職當月獎金	2.135	平均工作時數	-1.064
2	離職當月獎金變化	-1.555	離職當月獎金	0.877
3	每月平均獎金	-1.416	每月平均獎金	-0.841
4	是否轉正	-0.949	休假日工作比例	-0.710
5	平均工作時數	-0.807	離職當月獎金變化	-0.598
6	入職時雇用類型	0.778	入職月份	0.576

regular specification（enter進入方式）

看「非標準化係數」（estimate）跨年度比較在兩年皆排名前六重要的因素

因素	Year 1		Year 2	
	該年度 重要性排序	非標準化係數	該年度 重要性排序	非標準化係數
離職當月獎金	1	85.840	2	44.207
離職當月獎金變化	2	-7.864	5	-3.797
每月平均獎金	3	-52.953	3	-41.187
平均工作時數	5	-0.795	1	-1.091

stepwise specification

綜合考量「被放進model順序」和「標準化係數」評估個別因素重要性

被放進model順序	Year 1			Year 2		
	因素	標準化係數	依標準化係數排序	因素	標準化係數	依標準化係數排序
1	每月平均獎金	-1.471	3	每月平均獎金	-0.978	1
2	離職當月獎金	2.109	1	入職月份	0.435	5
3	離職當月獎金變化	-1.557	2	平均工作時數	-0.904	2
4	是否轉正	-0.994	4	入職時雇用類型	0.347	6
5	入職時雇用類型	0.840	5	離職當月獎金	0.795	3
6	平均工作時數	-0.773	6	離職當月獎金變化	-0.532	4

stepwise specification

看「非標準化係數」(estimate)跨年度比較在兩年皆排名前六重要的因素

因素	非標準化係數	
	Year 1	Year 2
每月平均獎金	-55.001	-47.846
離職當月獎金	84.786	40.076
離職當月獎金變化	-7.876	-3.380
平均工作時數	-0.761	-0.927
入職時雇用類型	2.324	1.166

比較進入方式和跨年度資料

- 共通點:
  - 找出的前六重要因素大致相同
  - 與獎金、工作時數有關的因素最值得關注，雇用類型、入職月份、是否轉正等也可能有影響
- 相異點:
  - 重要性排序略有不同

研究結果、建議與行動計畫

以下將分成三點說明，包含前頁比較中找出的兩大重要因素與其他分析方式建議：

- 工作時數
- 獎金
- 其他可採取的分析方式

工作時數

- 係數為負(離職編碼為1)→工作時數越長，越不離職
- 推測:工作時數越長，代表越有加班機會，財務報酬較高，較願意留任
  - 工作時數和每月平均薪資、最後一個月薪資在兩年皆有中度左右正相關

Year 1

Variable		SalaryBonusMean	SalaryMean	LastMonthSalaryBonus	LastMonthSalary	LastMonthSalaryPct	LastMonthSalaryBonusPct	WorkHourMean
1. SalaryBonusMean	Pearson's r	—						
2. SalaryMean	Pearson's r	0.643***	—					
3. LastMonthSalaryBonus	Pearson's r	0.573***	0.362***	—				
4. LastMonthSalary	Pearson's r	0.297***	0.600***	0.542***	—			
5. LastMonthSalaryPct	Pearson's r	−0.063***	−0.070***	0.033***	0.168***	—		
6. LastMonthSalaryBonusPct	Pearson's r	−0.364***	−0.201***	0.327***	0.326***	0.183***	—	
7. WorkHourMean	Pearson's r	0.251***	0.506***	0.162***	0.377***	0.013	−0.011	—

* p < .05, ** p < .01, *** p < .001

Year 2

Variable		SalaryBonusMean	SalaryMean	LastMonthSalaryBonus	LastMonthSalary	LastMonthSalaryPct	LastMonthSalaryBonusPct	WorkHourMean
1. SalaryBonusMean	Pearson's r	—						
2. SalaryMean	Pearson's r	0.623***	—					
3. LastMonthSalaryBonus	Pearson's r	0.511***	0.367***	—				
4. LastMonthSalary	Pearson's r	0.343***	0.604***	0.625***	—			
5. LastMonthSalaryPct	Pearson's r	−0.009	−0.005	0.046***	0.110***	—		
6. LastMonthSalaryBonusPct	Pearson's r	−0.176***	−0.080***	0.444***	0.356***	0.091***	—	
7. WorkHourMean	Pearson's r	0.210***	0.385***	0.132***	0.288***	0.028**	−0.020*	—

* p < .05, ** p < .01, *** p < .001

獎金

- 「每月平均獎金」及「離職當月獎金變化」係數為負，「離職當月獎金」係數為正

  →「每月平均獎金」、「離職當月獎金變化」相較前月越高，越不離職；「離職當月獎金」愈高，愈可能離職

-

建議行動計畫

★　數據分析發現工時與獎金是影響員工是否留任的主要關鍵

HR流程	招募	薪資設計
現況	● 在相對薪資水準較低的5-6月招募人員 ● 出現急單時提高仲介廠商招募傭金 ● 人力招聘流程無法有效呼應訂單調整，進而造成閒置人力	● 透過留任獎金，留住原有的生 產線人員 ● 留任獎金級距依統計分析的結果，調整合理的級距與金額
建議調整方向	1. 改良人力招聘流程，減少新進人員閒置的情形而提升整體人員工時 2. 招募時強調公司內部獎金制度	落實激勵性薪資設計，並將制度透明化，促使員工為積極達成目標而持續投入工作
行動	● 與人力仲介公司協調如何提升招募彈性 ● 在招募活動強調吸引人的獎金制度	調查目前的金錢獎酬中，何種獎金對員工留任影響最大、留任獎金的規劃是否公平、透明、有效益，有必要則改善薪酬制度

其他可採取的分析方式

本組認為，亦可以考慮使用machine learning，如Ramdon forest、boosting等技術建立預測模型，原因如下：

(1) 本次分析目的在於挑選出重要的factor

(2) 機器學習對於變量的限制條件少，就算是類別變項也ok

Year 1 - Boosting

Boosting Classification

Boosting Classification

Trees	Shrinkage	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy
100	0.10000	6056	1515	1892	0.99340	0.99260

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

Data Split

Confusion Matrix

		Predicted	
		0	1
Observed	0	20	8
	1	6	1858

- 「獎金」亦為最重要變數
- 「是否轉正」的相對重要性提高

Relative Influence

	Relative Influence
SalaryBonusMean	79.13542
Turn	8.02313
LastMonthSalaryBonus	7.27694
LastMonthSalaryPct	4.14270
WorkOnRestDay	0.53794
LeavePeakSeason	0.31542
WorkHourMean	0.29585
OriRecruitCat	0.20669
ShiftCateoryMost	0.06590
Gender	0.00000
LastMonthSalaryBonusPct	0.00000
ShiftRatio	0.00000
AgeRange	0.00000
AreaName	0.00000
OnBoardMonth	0.00000
DormAreaMost	0.00000

Year 1 - Random Forest

Random Forest Classification

Random Forest Classification

Trees	Features per split	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy	OOB Accuracy
78	4	6056	1515	1892	0.99538	0.99683	0.99730

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

Data Split

Confusion Matrix

		Predicted	
		0	1
Observed	0	28	0
	1	6	1858

Feature Importance ▼

	Mean decrease in accuracy	Total increase in node purity
SalaryBonusMean	0.09642	0.01309
LastMonthSalaryPct	0.26952	0.01024
LastMonthSalaryBonus	0.35343	0.00895
LastMonthSalaryBonusPct	0.09153	0.00502
Turn	0.05453	0.00487
OnBoardMonth	0.00853	0.00112
WorkOnRestDay	0.15859	0.00059
LeavePeakSeason	0.16431	0.00038
WorkHourMean	0.00526	0.00024
DormAreaMost	−0.00221	0.00001
ShiftRatio	0.11562	0.00001
ShiftCateoryMost	0.01228	7.03973×10^{-6}
AreaName	−0.00471	-9.39699×10^{-8}
AgeRange	0.00197	−0.00001
Gender	0.00078	−0.00004
OriRecruitCat	−0.00036	−0.00015

- 「獎金」亦為最重要變數
- 「是否轉正」的相對重要性提高

Year 2 - Boosting

Boosting Classification ▼

Boosting Classification

Trees	Shrinkage	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy
100	0.10000	7472	1868	2335	0.96734	0.96788

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

Data Split

Confusion Matrix

		Predicted	
		0	1
Observed	0	130	41
	1	34	2130

- 「獎金」亦為最重要變數
- 「休假日工作比例」、「入職月份」的相對重要性提高

Relative Influence ▼

	Relative Influence
SalaryBonusMean	42.09650
LastMonthSalaryBonus	23.04679
WorkOnRestDay	17.25465
OnBoardMonth	10.03272
OriRecruitCat	2.80116
LeavePeakSeason	2.62712
WorkHourMean	0.81938
LastMonthSalaryPct	0.47788
ShiftRatio	0.46765
ShiftCategoryMost	0.25894
DormAreaMost	0.06861
AgeRange	0.04858
Gender	0.00000
Turn	0.00000
LastMonthSalaryBonusPct	0.00000
AreaName	0.00000

Year 2 - Random Forest

Random Forest Classification ▼

Random Forest Classification ▼

Trees	Features per split	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy	OOB Accuracy
45	4	7472	1868	2335	0.97056	0.97388	0.98280

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

Data Split

Confusion Matrix

		Predicted	
		0	1
Observed	0	143	28
	1	33	2131

Feature Importance ▼

	Mean decrease in accuracy	Total increase in node purity
WorkOnRestDay	0.29799	0.01330
LastMonthSalaryBonus	0.29011	0.01328
SalaryBonusMean	0.04722	0.01076
LastMonthSalaryPct	0.12471	0.00411
OnBoardMonth	0.06575	0.00226
LastMonthSalaryBonusPct	0.07264	0.00197
Turn	0.01264	0.00100
ShiftCategoryMost	0.00539	0.00058
AgeRange	0.00232	0.00027
DormAreaMost	−0.00376	0.00021
WorkHourMean	0.03492	0.00003
OriRecruitCat	0.01041	−0.00005
Gender	0.00012	−0.00006
ShiftRatio	0.07502	−0.00016
AreaName	−0.00225	−0.00016
LeavePeakSeason	0.13477	−0.00081

- 「獎金」亦為最重要變數
- 「休假日工作比例」、「入職月份」的相對重要性提高