



王道銀行

所得代理模型專案 模型開發報告書_模型開發篇

版本 4.0

免責聲明

本報告所載資料是為王道銀行 0001438012 個人信用貸款所得代理模型開發專案編撰。任何其他人士不得依賴本報告作任何其他用途，本所/本公司概不就此對任何其他人士承擔任何責任。未經本所/本公司書面同意，任何人士不得傳送、揭露、引用或引述本報告的全部或部分內容。

The information contained herein has been prepared for O-Bank Co., Ltd. pursuant to Engagement Letter No. 0001438012, Income proxy model project. It should not be relied upon by any other party for any other purpose and we expressly disclaim any liability or duty to any other party in this respect. It should not be disclosed, referred to or quoted in whole or in part without our prior written consent.



修訂紀錄 (Modification Record)

版次	作者	摘要	發行/修訂生效日期
V1.0	KPMG	初稿	2023.09.30
V2.0	KPMG	補充第 3、4 章內容，並新增第 5 章附件	2023.10.13
V2.1	KPMG	修正 4.6 章節內容	2023.10.16
V3.0	KPMG	4.6~4.12 各節的模型效度小節追加「新增測試集」的效度結果呈現	2023.12.11
V4.0	KPMG	修正 4.11 與 4.12 節的模型結果與效度，並新增 5.7 節附件	2023.12.25

目錄

1.	前言與摘要	7
1.1	專案背景與範圍	7
1.2	模型摘要	8
2.	模型開發方法論	9
2.1	總體趨勢子模型	9
2.1.1	模型開發架構與流程	9
2.1.2	模型開發方法說明	10
2.2	個體差異子模型	12
2.2.1	模型開發架構與流程	12
2.2.2	多元線性迴歸	13
2.2.3	分量迴歸	14
2.2.4	模型開發具體步驟	14
3.	總體趨勢子模型開發	17
3.1	模型資料處理	17
3.1.1	主計處薪情平臺	17
3.1.2	軍公教警人事管轄機關之公開資訊	18
3.2	模型客群分類	21
3.3	建立總體趨勢薪資彙總表	28
3.4	合併總體趨勢薪資彙總表與內部資料	29
3.5	職類別薪資資料處理	34
4.	個體差異子模型開發	37
4.1	樣本排除條件	37
4.2	應變數定義與分析	40
4.2.1	應變數定義	40
4.3	長清單定義	41
4.4	抽樣設計與結果	41
4.5	單變量分析結果	42
4.6	一般服務業(C10)子模型多變量分析與模型驗證結果	43
4.6.1	模型設定	43
4.6.2	模型結果	43
4.6.3	模型效度	45

4.6.4	模型變數相關說明	47
4.6.5	分量迴歸模型結果	51
4.6.6	拔靴法交叉驗證	54
4.7	工程及批發零售業(C20)子模型多變量分析與模型驗證結果	56
4.7.1	模型設定	56
4.7.2	模型結果	56
4.7.3	模型效度	58
4.7.4	模型變數相關說明	60
4.7.5	分量迴歸模型結果	65
4.7.6	拔靴法交叉驗證	68
4.8	技術勞動業(C30)子模型多變量分析與模型驗證結果	70
4.8.1	模型設定	70
4.8.2	模型結果	70
4.8.3	模型效度	72
4.8.4	模型變數相關說明	74
4.8.5	分量迴歸模型結果	80
4.8.6	拔靴法交叉驗證	84
4.9	專業服務業(C40)子模型多變量分析與模型驗證結果	86
4.9.1	模型設定	86
4.9.2	模型結果	86
4.9.3	模型效度	88
4.9.4	模型變數相關說明	90
4.9.5	分量迴歸模型結果	93
4.9.6	拔靴法交叉驗證	96
4.10	資通金融能源業(C50)子模型多變量分析與模型驗證結果	98
4.10.1	模型設定	98
4.10.2	模型結果	98
4.10.3	模型效度	100
4.10.4	模型變數相關說明	102
4.10.5	分量迴歸模型結果	109
4.10.6	拔靴法交叉驗證	112
4.11	軍公教警類(C60)子模型多變量分析與模型驗證結果	114
4.11.1	模型設定	114
4.11.2	模型結果	114
4.11.3	模型效度	116
4.11.4	模型變數相關說明	118
4.11.5	分量迴歸模型結果	121
4.11.6	拔靴法交叉驗證	124

4.12	非穩收族群子模型多變量分析與模型驗證結果.....	126
4.12.1	模型設定	126
4.12.2	模型結果	126
4.12.3	模型效度	128
4.12.4	模型變數相關說明	130
4.12.5	分量迴歸模型結果	136
4.12.6	拔靴法交叉驗證.....	139
5.	附錄	142
5.1	行業與職稱別對照表	142
5.2	修訂行業別對照表	142
5.3	勞動部職業名稱及工作說明.....	142
5.4	109 年各縣市工業及服務業企業僱用員工薪資統計說明.....	142
5.5	110 年工業及服務業受僱員工全年總薪資中位數及分布統計結果	142
5.6	特徵變數轉換檔	143
5.7	模型效度一覽	143

1. 前言與摘要

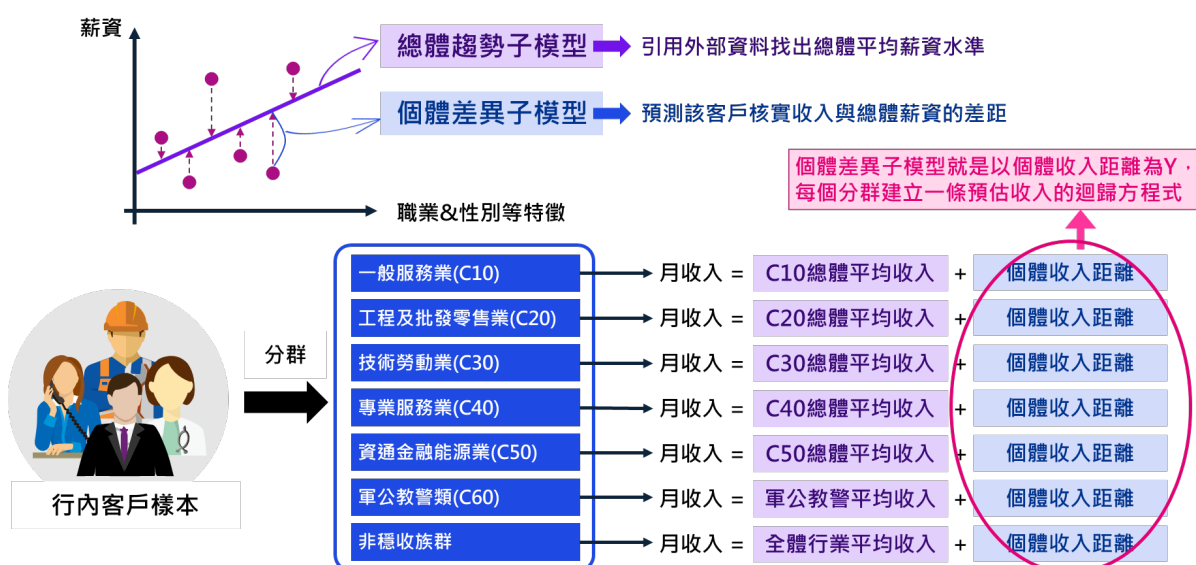
1.1 專案背景與範圍

王道商業銀行（以下簡稱王道銀行），欲藉由所得代理模型（Income Proxy Model）有效針對客戶所得進行評估，進而取代傳統方式徵提實際收入文件的程序，以加深客戶申辦時的良好體驗，進而增加進件率。因此委請安侯企業管理股份有限公司（以下簡稱KPMG），進行相關的模型建置作業。

本模型由「總體趨勢子模型」與「個體差異子模型」組合而成。總體趨勢子模型引用主計處薪情平臺、軍公教警待遇與人數統計等外部資料，給定該行業類群的總體平均薪資水準。個體差異子模型則是根據該客戶之特徵，預測該客戶核實收入與總體薪資之間的收入距離，故結合總體趨勢子模型與個體差異子模型即可有效推估客戶收入。

為了使模型充分反映不同行業下的族群特性，KPMG 根據行內資料將模型分群為 7 類，如下表所示：

序號	模型分類	模型分類名稱
1	C10	一般服務業
2	C20	工程及批發零售業
3	C30	技術勞動業
4	C40	專業服務業
5	C50	資通金融能源業
6	C60	軍公教警類
7	非穩收族群	非穩收族群



本文件將詳述以上 7 類模型之總體趨勢子模型與個體差異子模型，內容包含：開發方法、流程步驟、開發結果以及驗證結果。

1.2 模型摘要

KPMG 於職位價值模型開發最終產出共 7 個模型，模型開發結果不論因子選取、效度上都符合業務直覺與水準，同時經過交叉驗證後模型效度與模型係數均達到穩定，可做為王道銀行業務使用，茲摘要如下：

模型變數&MAPE	C10	C20	C30	C40	C50	C60	非穩收族群
職稱倍率	V	V	V	V	V		
年齡						V	V
性別倍率/ 性別年薪中位數	V	V	V	V	V		V
推估總年資/ 推估總年資註記	V	V	V		V		
工作地年薪平均數	V	V	V		V		
戶籍地年薪 中位數比率							V
教育程度年薪 中位數		V	V	V		V	V
上市櫃公司溢酬			V		V		V
行業別虛擬變數/ 客群區分註記/ 外商金融機構註記				V	V	V	V
離島服務註記						V	
近 6/12 個月平均每 發卡銀行永久額度	V	V	V	V	V	V	V
無擔保貸款總額度	V	V	V	V	V	V	V
近 6/12 個月高消費 註記	V	V	V		V		
有擔保貸款高額度 註記	V	V	V	V	V	V	V
信用卡一年期 總消費金額						V	V
訓練集 MAPE	21.61%	24.77%	24.64%	27.97%	23.37%	16.40%	33.22%
測試集 MAPE	22.55%	24.82%	24.41%	28.19%	24.37%	16.13%	32.81%
全樣本 MAPE	21.80%	24.78%	24.59%	28.01%	23.57%	16.35%	33.14%

2. 模型開發方法論

本專案在模型建置的過程中，將區分總體趨勢子模型與個體差異子模型，模型開發流程步驟如下：



2.1 總體趨勢子模型

2.1.1 模型開發架構與流程

總體趨勢子模型開發步驟與流程如下說明。

- 步驟一：理解所蒐集之資料源定義、變數與使用方式，詳細內容說明請參考「王道銀行所得代理模型專案 模型開發報告書_資料處理篇」。
- 步驟二：參考前述報告書第二章之說明，不同來源之外部資料皆需透過相關手動彙整將資料轉換成能夠匯入分析軟體進行分析的格式，後續再依據各項資料處理邏輯完成各產業別之內外部行業別對照表、總體趨勢模型薪資對照表等。
- 步驟三：利用統計分析包含決策樹、K Means 等演算法與專家調整精簡客群類別，考量產業特性、薪資結構與樣本分布等面向將相似的族群歸於同一類別，作為個體差異子模型分類之基礎。



2.1.2 模型開發方法說明

總體趨勢子模型除針對薪資資料進行彙整外，其中在行業別分類上，透過決策樹與 K Means 搭配運用並結合專家調整完成最終之行業別分類，後續再透過單因子變異數分析檢視各組間的差異程度，以下針對三種統計方式進行說明。

■ 決策樹

決策樹(Decision Tree)是一種透過樹狀分枝的概念作為決策的模式，其主要是利用遞迴劃分(recursive partitioning)進行特徵選取而得的。遞迴劃分就是將所有的觀察樣本都放入一個大箱子，再將每個資料進行反覆劃分的過程。在使用決策樹演算法進行分析之前，首要工作就是選擇適當的演算法。一般來說，演算法的應用為根據所要分析資料的特性以及型態等進行選擇。而本次所使用之決策樹演算法說明如下：

CART(Classification and Regression Trees)

CART(分類迴歸樹)演算法為 Brieman et al.於 1984 年所發表出來。CART 藉由一個單一輸入變數函數，在每一個節點分個資料，以建構一個二分式決策樹。因此，第一個任務是決定哪一個自變數可成為最好的分隔變數。此處所謂最好分隔的定義是能夠將資料最完善的分配到一個單一類別支配的群體。如同其名，CART 演算法既可以用於創建分類樹(Classification Tree)，也可以用於創建迴歸樹(Regression Tree)，兩者在建樹的過程稍有差異。

本次欲分類的觀察值為各行業別之薪資，因此觀察值取值是連續的、沒有分類標籤，只有根據觀察資料得出的值來創建一個預測的規則。在這種情況下，Classification Tree 的最優劃分規則就無能為力，CART 會使用最小平方誤差(Squared Residuals Minimization)來決定 Regression Tree 的最優劃分，該劃分準則是期望劃分之後的子樹誤差方差最小。以下針對最小平方誤差說明：

假設訓練資料 D 為 $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ ，輸出 Y 為連續變量，將輸入劃分為 M 個區域，分別為 R_1, R_2, \dots, R_M ，每個區域的輸出值分別為： C_1, C_2, \dots, C_m ，則迴歸樹模型可表示為：

$$f(x) = \sum_{i=1}^m C_i I(x \in R_i)$$

則平方誤差為

$$\sum_{x_i \in R_i}^m (y - f(x_i))^2$$

假如使用屬性 j 的取值 s 來將輸入空間劃分為兩個區域，分別為：

$$R_1(j, s) = \{x | x^{(j)} \leq s\} \text{ 和 } R_2(j, s) = \{x | x^{(j)} > s\}$$

我們需要最小化損失函數為：

$$\min_{j, s} \left[\sum_{x_i \in R_1(j, s)} (y_i - c_1)^2 + \sum_{x_i \in R_2(j, s)} (y_i - c_2)^2 \right]$$

其中 c_1, c_2 分別為 R_1, R_2 區間內的輸出平均值。

為了使平方誤差最小，我們需要依次對每個特徵的每個取值進行評估，計算出當前每一個可能的切分點的誤差，最後選擇切分誤差最小的點將輸入空間切分為兩個部分，然後遞迴上述步驟，直到切分結束。

■ K Means 演算法

K Means 為一種聚類分析方法，其目的主要是把 n 個點劃分到 K 個集合中，使得每個點和各自群心的誤差平方和越小越好，最小化群內資料。**K Means** 演算法說明如下：

假設有一組資料 (x_1, x_2, \dots, x_n) ，其中每個觀測值皆滿足 $x_i \in R^d$ ，**K Means** 主要是把 n 個觀測值劃分到 K 個集合中 ($K \leq n$)，使得組內平方和最小，數學公式如下：

$$\arg \min_S = \sum_{i=1}^k \sum_{X \in S_i} \|X - \mu_i\|^2$$

其中 S_i 為各集合； μ_i 是 S_i 中所有點的均值； $\|X - \mu_i\|$ 為歐式距離

■ 變異數分析(ANOVA)

變異數分析(ANOVA)用於比較組與組之間對於一個應變數的「平均數」是否有顯著差異。變異數分析之檢定假設與公式說明如下：

虛無假設(H_0)：各組平均數皆相等 V.S 對立假設(H_1)：至少有兩組間平均數不相等

	平方和	均方	F 檢定	p-value
組間	SS_G	MS_G	MS_G/MS_W	查表
組內	SS_W	MS_W		
全體	SS_T			

$$\text{總變異} : SS_T = \sum \sum (y_{ij} - \bar{y})^2$$

$$\text{組間變異：} SS_G = \sum n_i (y_i - \bar{y})^2$$

$$\text{組內變異：} SS_W = \sum \sum (y_{ij} - \bar{y}_i)^2$$

$$MS_G = \frac{SS_G}{k-1} ; MS_W = \frac{SS_W}{n-k}$$

其中 k 為組數；n 為樣本數

2.2 個體差異子模型

2.2.1 模型開發架構與流程

「個體差異子模型」的建置，主要是以主計處行業別能與行內資料進行串接且為受雇員工之樣本為基準，進行客戶所得的預測。由於不同行業有不同的薪資水準，本模型透過總體趨勢子模型階段先將薪資結構相近之行業別進行分類，後續於個體差異子模型階段根據行業別分類分別進行模型開發。其作法能夠針對不同客群之特徵，分別納入與該客群攸關的解釋變數，並分別進行迴歸模型的配適，以避免這些會因不同分類而有不同特徵的解釋變數在整體樣本下解釋力因未分類原因而下降，且隨著不同分類分別進行配適也有提升預測效度的優勢，冀以提升整體效度的同時也符合業務應用的直觀。

個體差異子模型開發步驟與流程如下說明。

- 定義變數與樣本範圍：預測的目標變數為客戶核實薪資與總體趨勢子模型的收入距離，而解釋變數則需要依據資料的可得性以及相關的業務經驗，在資料整理階段備妥相關的變數長清單，以作為模型潛在自變數的選擇。最後，要確認建模樣本範圍，排除極端的離群值以避免影響模型配適的效果。
- 單變量分析：分別針對七個子模型的變數長清單計算各項統計指標、進行相關性分析，以及利用單因子迴歸計算 MAPE，並根據這些敘述統計資訊來挑選出進入多變量分析的短清單變數。
- 多變量分析：而進入多變量分析的環節後，個體差異子模型以多元線性迴歸作為模型擬合主要使用的統計方法，同時針對個體差異子模型估計了第 40 百分位數(Q40)、第 35 百分位數(Q35)、第 30 百分位數(Q30)、第 25 百分位數(Q25)、第 20 百分位數(Q20)的分量迴歸模型，其目的是取得保守性估計的結果以搭配後續的策略應用，協助業管單位在風險可控的前提下追求合理的目標利潤。其中有關多元線性迴歸與分量迴歸的方法，詳見「2.2.2 多元線性迴歸」與「2.2.3 分量迴歸」二節之說明。



2.2.2 多元線性迴歸

多元線性迴歸(Multiple linear regression)為一種統計的預測模型，以線性方程式來描述兩個以上的自變數(x)與一個應變數(y)之間的關係，可以藉由普通最小平方法(OLS)來求解出多元線性迴歸模型的模型係數，該模型係數即代表自變數對於應變數的偏效果，也就是自變數的變動對應變數所造成的影響，故藉由建立多元線性迴歸模型即可根據已知的自變數值，對未知的應變數進行推估。

多元線性迴歸模型一般的模型假設如下：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

其中， β_0 代表截距項， β_1, \dots, β_n 代表模型係數， ε 代表誤差項。

以普通最小平方法進行估計時，則需令誤差平方和(Sum of the Squared Errors, SSE)函數為：

$$f(e) = \sum_{i=1}^n \varepsilon^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_i x_i)^2$$

對上列函數的所有模型係數分別進行偏微分後，即可得到有 $n+1$ 條方程的方程組，對該方程組進行求解，即可得到當誤差平方和為最小時模型係數的估計結果。

而當應變數與自變數之間的變化關係是以比例的方式存在時，可以採用自然對數轉換來修正模型設定，例如：

1. 當 X 平均每上升 1%，Y 約增加 b%

表示應變數與自變數兩邊皆呈現比例的成長，此時可以考慮將應變數與自變數同時取自然對數如下：

$$\log(Y) = a + b \log(X) + \cdots$$

2. 當 X 平均增加 1 單位，Y 約增加 b%

表示自變數固定數值的增加會導致應變數呈現比例的成長，此時可以考慮將自變數維持不變下，讓應變數取自然對數如下：

$$\log(Y) = a + b X + \dots$$

2.2.3 分量迴歸

相較於典型的迴歸分析研究自變數與應變數的條件期望之間的關係，分量迴歸(Quantile regression)則是研究自變數與應變數的在不同分位數的條件關係，藉此能夠更清楚地描繪出應變數整體的分配特性。利用分量迴歸方法，對不同分位數的應變數進行擬合，即可得到相較多元迴歸模型更加保守的估計結果。

分量迴歸模型一般的模型假設如下：

$$Q_p(y) = \beta_0(p) + \beta_1(p)x_1 + \beta_2(p)x_2 + \dots + \beta_n(p)x_n + \varepsilon$$

其中， $\beta_0(p), \beta_1(p), \dots, \beta_n(p)$ 代表模型係數， ε 代表誤差項。

此時的模型係數並非單一的估計值，而是隨著估計的分位數 p 的不同而跟著變動的值，故以函數的型態表示之。

而分量迴歸求解的方式也與多元線性迴歸的最小平方法不同，分量迴歸是採取極小化加權離差絕對值總和(Least Absolute Deviation, LAD)，也就是在不同分量位置給予不同的權數，並以線性規劃方式求解。

2.2.4 模型開發具體步驟

■ 單變量分析(Single Factor Analysis)

單變量分析之目的是為了從全部變數中篩選出較具檢定力與代表性的變數，再進行後續的多變量分析與模型開發。單變量分析是針對單一變數進行敘述統計，並綜合考量相關性分析與單因子迴歸 MAPE 之排序，從不同角度進行變數之篩選，確保在後續分析中所使用的變數在理論與實務上具有解釋能力與預測效果。

單變量分析步驟：

1. 檢視樣本敘述統計值，包含缺失值與零值的比例、最大值與最小值、唯一值筆數、眾數、四分位數、第 1 與第 99 百分位數、平均值、標準差、離群值、正值與負值筆數等，確定樣本篩選範圍與排除條件。
2. 計算相關係數、MAPE，先後以 MAPE 與相關係數進行排序，並考量缺失值或零值比例在合理情況下，優先挑選 MAPE 較低、相關係數較高的變數作為短清單變數。

■ 多變量分析(Multiple Factors Analysis)

選擇合適的變數是建立模型重要的步驟，以單變量分析時的 MAPE 排序的結果當作變數投入順序的參考是一種可行的方式。

特別強調的是，多變量分析初期選擇投入變數時可供參考的，除了單變量分析時的 MAPE 排序之外，尚須考慮以下相關變數選取的原則，反覆將變數投入多元迴歸中試行而取得最終模型：

變數的遞增遞減趨勢需與常理及業務直覺相符。

意義相似之變數(例如最近 3、6、12 個月平均每發卡銀行永久額度)，則挑選 MAPE 較低者作為代表。

變數顯著性檢定結果皆為顯著，即 t 檢定之 P-value 應小於 0.1。

確認變數無共線性，即變異數膨脹因素(Variance Inflation Faction, VIF)需小於 10。

VIF 值愈小代表越沒有共線性問題，公式如下：

$$VIF_j = \frac{1}{1 - R_j^2}, j = 1, 2, \dots, p$$

其中 $(1-R_j^2)$ ，稱做為容忍度。若 $VIF_j > 10$ ，則表示第 j 個預測變數為共線性變數。

變數進入能夠提升整體模型的效度，即 MAPE 能夠下降。

變數在交叉驗證時表現穩定。

■ 模型效度指標(Model Performance Index)

平均百分比誤差(Mean Absolute Percentage Error, MAPE)用於評估模型預測的準確度，其使用了誤差百分比來量測預測值與實際值之間的差距，而非誤差的絕對數值，比較不容易被極值所影響，且誤差百分比取絕對值，才不會造成正離差與負離差相互抵消而低估誤差，其公式如下：

$$MAPE = \frac{1}{n} \cdot \sum_i^n \left| \frac{(\text{預測值}_i - \text{實際值}_i)}{\text{實際值}_i} \right| \times 100\%$$

MAPE 的好處在於提供了一個容易比較的 benchmark，當預測值與實際值完全相同的時候，MAPE 值最小為 0，因此 MAPE 的結果越接近於 0 越好，但是當實際值很小的時候，一點點的誤差就容易導致 MAPE 的膨脹，因此在核實月收入較低的區間，MAPE 容易受到分母較低的影響而放大了估計誤差呈現的差異效果。

MAPE 隨著模型種類的不同有不同的效度門檻，通常不建議類比其他同類模型的門檻值，建議是在可得資料的與樣本之下追求最低的 **MAPE** 值，再加以監督與評估其變化，同時搭配應用策略以進行模型的運用。

■ 模型驗證(Model Validation)

樣本內/樣本外驗證

在個體差異子模型建置過程中，為了能夠有效評估模型之預測能力以及診斷是否允當，必須透過驗證來檢視成效。而樣本內與樣本外的測試，通常是建模時抽取較大比例(如 80%)的樣本做為訓練集，留下較小比例(如 20%)的樣本進行套用(測試集)。便是利用樣本內的客戶(訓練集)進行模型的擬合，再套用至樣本外的客戶(測試集)，觀察兩不同樣本的效度，其目的是來測試由樣本內的客戶所訓練出來的模型，是否套用在樣本外的客戶在效度上沒有減損或差異過大，以致於有過度擬合(**Overfitting**)的問題發生。

拔靴法(Bootstrapping)交叉驗證

當樣本稀少與難以蒐集時，一般會採用 **Bootstrapping** 交叉驗證來評估模型。交叉驗證就是重複的對同一組數據進行抽樣，重新組合為不同的訓練集和測試集，因此於某次訓練集中的樣本在下次亦可能成為測試集中的樣本。以下為 **Bootstrapping** 交叉驗證進行步驟：

1. 採隨機抽出各年收入區間 80%與 20%的樣本。收入區間定義與樣本內外測試執行時相同。
2. 將步驟 1 得到的 80%與 20%的樣本套用到欲測試的模型，紀錄其效度指標。
3. 將 80%的樣本以模型中原有的變數重新訓練，並紀錄新的模型係數。
4. 重複步驟 1 到 3 共 200 次。
5. 計算模型係數和效度指標在這 200 次隨機抽樣中的的平均數、標準差及變異係數(**Coefficient of Variance**)。
6. 變異係數之絕對值須小於 30%，代表抽樣後樣本偏離不大。

3. 總體趨勢子模型開發

依據 2.1 章節之總體趨勢子模型開發方法說明，總體趨勢子模型開發包含模型資料處理、模型客群分類以及最終分類之總體趨勢對照表開發，以下針對各步驟進行說明：

3.1 模型資料處理

在總體趨勢子模型之資料整理方面，須先針對不同來源之資料進行資料型別轉換與資料欄位衍生等相關資料處理，不同來源相對也會有不同的資料處理方式，以下將拆分各來源資料進行說明：

3.1.1 主計處薪情平臺

主計處薪情平台在資料年月欄位的儲存上為民國年，因此須將資料年月型態轉成為西元年與行內資料年月型態一致，後續才能進行內外部資料串接。此外，薪情平台資料是以月為單位，資料區間為 99 年 8 月至 111 年 12 月。從資料統計值定義來看，在每人每月總薪資包含了經常性與非經常性薪資，因此在薪資資料彙整上可以將 12 個月的統計薪資做加總計算年薪後再除以 12 去推算各行業的平均月薪資，比較到僅觀察單月份之總薪資，推算後之平均月薪資會更為精確。而在計算年薪的區間選擇上，考量到運用在相同範圍的勞動部職類別薪資是以 7 月為基準月進行統計，因此年薪計算區間為前一年度 8 月至當年度 7 月之十二個月月薪合計加總，後續透過程式開發執行各年度之年薪計算，總體趨勢薪資彙整以年為單位並以 7 月為基準月，在計算完各年度年薪後再除以 12 取得各年度 7 月份之平均月薪資，未達 12 個月年薪計算之樣本則予以排除。以 2020 年 7 月為例，將薪情平臺薪資彙整表¹依不同行業別從 2019 年 8 月份至 2020 年 7 月份的總薪資包含總薪資_全體(INCOME_ALL)、總薪資_男(INCOME_M)與總薪資_女(INCOME_F)三個欄位分別作加總取得年薪再除以 12 作為 2020 年 7 月的平均月薪包含平均月薪_全體(MONTH_INCOME_ALL)、平均月薪_男(MONTH_INCOME_M)與平均月薪_女(MONTH_INCOME_F)。其他欄位如資料年月、行業人數(含全體、男、女)等皆以各年度 7 月資料為基準，加權月薪欄位(含全體、男、女)則利用上述已完成計算平均月薪乘以行業人數求得(全體、男、女分別計算)。依據前述之邏輯最終輸出資料表

¹ 薪情平臺薪資彙整表詳見「王道銀行所得代理模型專案 模型開發報告書_資料處理篇」之 2.1.3 節資料欄位定義與處理

名為 TABLE_SALARY_BASE，該資料表區間為 2011 年至 2022 年並以 7 月為基準月共 17 個行業別，204 筆樣本點。相關資料表之薪資資料彙整欄位及統計值說明如下：

序號	欄位名稱	欄位說明	欄位註解
1	CAREER_CODE	行業代碼大類	
2	CAREER_CH	行業名稱大類	
3	DATE_YM	原始資料年月	民國年月
4	DATA_YM	資料年月	轉換後欄位，為西元年月
5	YEAR	資料年	衍生欄位，作為模型開發使用
6	MONTH	資料月	衍生欄位，作為模型開發使用
7	NUM_PPL_ALL	行業人數_全體	以各年度 7 月為基準
8	NUM_PPL_M	行業人數_男	以各年度 7 月為基準
9	NUM_PPL_F	行業人數_女	以各年度 7 月為基準
10	YEAR_INCOME_ALL	年薪_全體	加總全體人數之前一年度 8 月至當年度 7 月月薪資
11	YEAR_INCOME_M	年薪_男	加總男性之前一年度 8 月至當年度 7 月月薪資
12	YEAR_INCOME_F	年薪_女	加總女性之前一年度 8 月至當年度 7 月月薪資
13	MONTH_INCOME_ALL	平均月薪_全體	年薪_全體/12
14	MONTH_INCOME_M	平均月薪_男	年薪_男/12
15	MONTH_INCOME_F	平均月薪_女	年薪_女/12
16	MONTH_INCOME_PPL_ALL	加權月薪_全體	平均月薪_全體*行業人數_全體，作為後續子模型分類計算使用
17	MONTH_INCOME_PPL_M	加權月薪_男	平均月薪_男*行業人數_男，作為後續子模型分類計算使用
18	MONTH_INCOME_PPL_F	加權月薪_女	平均月薪_女*行業人數_女，作為後續子模型分類計算使用

3.1.2 軍公教警人事管轄機關之公開資訊

軍公教警類別之外部薪資資料上，因該類別須透過年齡區間進行資料串接，年齡上界 [AGE_TOP] 之資料型態須轉換成數值型態，才能供後續內外部資料進行合併。此外，依據目前的資料規格僅須再推算平均月薪資即可完成該類別之對照表的建立，參考到行政院 110 年軍公教人員年終工作獎金發給注意事項的說明，符合發給條件會給予 1.5 個

月之年終工作獎金，以上述為基礎去推算年薪為月薪的 13.5 個月，因此在公教警各自的薪資對照表應先將男平均月薪資 (AVG_SALARY_M) 與女平均月薪資 (AVG_SALARY_F) 進行加權平均取得全體人數的平均月薪資 (AVG_SALARY_ALL)，再進行年薪與調整後之平均月薪資 (MONTH_SALARY_ALL_WA) 的計算，即 $AVG_SALARY_ALL * 13.5 / 12$ ，而軍人因無區分性別的月薪資，僅透過平均月薪資去推算年薪與調整後之平均月薪資 (MONTH_SALARY_ALL_WA)，即 $AVG_SALARY * 13.5 / 12$ ，完成後即可取得軍公教警各別之總體趨勢薪資彙總表。彙總表之欄位與統計值說明如下：

■ 公務人員薪資彙總表[DATA_PUBLIC]

序號	欄位名稱	欄位說明	欄位註解
1	DATA_YEAR	資料年度	
2	BGN_YM	資料套用起始年月	
3	END_YM	資料套用結束年月	
4	AGE_BOT	年齡下界	
5	AGE_TOP	年齡上界	
6	COUNT_M	男性總人數	
7	COUNT_F	女性總人數	
8	AVG_SALARY_M	男平均月薪資	原始資料薪資
9	AVG_SALARY_F	女平均月薪資	原始資料薪資
10	AVG_SALARY_ALL	全體平均月薪資	由男、女平均月薪資加權計算平均而得
11	MONTH_SALARY_ALL_WA	全體平均月薪資_調整	由全體平均月薪資*13.5/12
12	MONTH_SALARY_M_WA	男平均月薪資_調整	由男平均月薪資*13.5/12
13	MONTH_SALARY_F_WA	女平均月薪資_調整	由女平均月薪資*13.5/12

■ 教師薪資彙總表[DATA_TEACHER]

序號	欄位名稱	欄位說明	欄位註解
1	DATA_YEAR	資料年度	
2	BGN_YM	資料套用起始年月	
3	END_YM	資料套用結束年月	
4	AGE_BOT	年齡下界	
5	AGE_TOP	年齡上界	
6	COUNT_M	男性總人數	

序號	欄位名稱	欄位說明	欄位註解
7	COUNT_F	女性總人數	
8	AVG_SALARY_M	男平均月薪資	原始資料薪資
9	AVG_SALARY_F	女平均月薪資	原始資料薪資
10	AVG_SALARY_ALL	全體平均月薪資	由男、女平均月薪資加權計算平均而得
11	MONTH_SALARY_ALL_WA	全體平均月薪資_調整	由全體平均月薪資*13.5/12
12	MONTH_SALARY_M_WA	男平均月薪資_調整	由男平均月薪資*13.5/12
13	MONTH_SALARY_F_WA	女平均月薪資_調整	由女平均月薪資*13.5/12

■ 警察薪資彙總表[DATA_POLICE]

序號	欄位名稱	欄位說明	欄位註解
1	DATA_YEAR	資料年度	
2	BGN_YM	資料套用起始年月	
3	END_YM	資料套用結束年月	
4	AGE_BOT	年齡下界	
5	AGE_TOP	年齡上界	
6	COUNT_M	男性總人數	
7	COUNT_F	女性總人數	
8	AVG_SALARY_M	男平均月薪資	原始資料薪資
9	AVG_SALARY_F	女平均月薪資	原始資料薪資
10	AVG_SALARY_ALL	全體平均月薪資	由男、女平均月薪資加權計算平均而得
11	MONTH_SALARY_ALL_WA	全體平均月薪資_調整	由全體平均月薪資*13.5/12
12	MONTH_SALARY_M_WA	男平均月薪資_調整	由男平均月薪資*13.5/12
13	MONTH_SALARY_F_WA	女平均月薪資_調整	由女平均月薪資*13.5/12

■ 軍人薪資彙總表[DATA_MILITARY]

序號	欄位名稱	欄位說明	欄位註解
1	DATA_YEAR	資料年度	
2	BGN_YM	資料套用起始年月	
3	END_YM	資料套用結束年月	
4	ARMY_TYPE	軍種	
5	AGE_BOT	年齡下界	

序號	欄位名稱	欄位說明	欄位註解
6	AGE_TOP	年齡上界	
7	AVG_SALARY	平均月薪資	原始資料薪資
8	MONTH_SALARY_ALL_WA	平均月薪資_調整	由平均月薪資*13.5/12

3.2 模型客群分類

為了利用主計處薪情平臺與軍公教警的外部薪資統計資料，首先需要建立行業別分類(CC_CODE_REVIEW)與主計處行業代碼(CAREER_CODE)的對照關係，但考量到行內的分類規則下較難以直接與主計處行業代碼進行匹配，為避免分類不當而導致錯誤套用總體趨勢薪資，針對行內外職業別分類的對照需進行妥適處理，茲說明如下：

1. 部份情況下無法單獨以行業別分類(CC_CODE_REVIEW)決定對應的主計處行業別代碼(CAREER_CODE)，仍需搭配公司名稱(COMPANY)中的關鍵字以判斷合適的分類，故 KPMG 綜合考量現行分類下的樣本結果與主計處分類規則，將初步處理規則彙整於附錄 5.1 行業與職稱別對照表「行業 CC_Code」中的備註欄位
2. 針對分類規則難以明確的行業別分類(CC_CODE_REVIEW)則需以人工方式進行認定，故依照附錄 5.2 修訂行業別對照表中「修訂後對照表」內「需調整樣本」為 T 的進行 JOIN，即當公司名稱(COMPANY)相同時，修正為人工認定的主計處行業別代碼(CAREER_CODE)。另外有一部分樣本則載於「剩餘樣本之對照表」，需以 APPLNO 進行 JOIN
3. 最後，CC_CODE_REVIEW=18 中審查意見(credit_comment)中有出現關鍵字「教官」的樣本歸類為 O-1(軍人)

除了確立行業別分類(CC_CODE_REVIEW)與主計處行業代碼(CAREER_CODE)的對照關係以外，因應建模需求還需要進行佣金收入非穩收客群、小商號負責人客群的判斷，茲說明如下：

■ 判斷佣金收入之非穩收客群

根據 2023/7/20 會議討論結果，穩收/非穩收的定義並非以 貴行政策的穩收 flag 為判斷依據，而是根據下列條件進行認定，即符合下列條件者穩收註記(STABLE_FLAG)賦值為 N，其他則賦值為 Y：

CC_code	OC_code
34-自由業	

35-網路賣家	
	24-佣金收入專業人員(房仲、汽車等獎金制)
	25-佣金收入專業人員(金融)
	26-佣金收入專業人員(高風險之合法登記娛樂場所、休閒場所、其他服務業-殯葬業)
	28-網路賣家、論件/按日計酬無固定薪人員
	29- 自由業(含論件/按日計酬無固定薪人員)
	32-自由業 (藝術、演藝事業等)
	33-自由業 (清潔工、學徒、搬運工、建築工人、快遞員、收費員、工讀生)
	34-其他
	47-佣金收入專業人員(一般公司、專櫃、門市)
	49-健身教練
	50-導遊、領隊
	61- 佣金收入專業人員
	66- 不動產經紀人

惟部分穩收註記(STABLE_FLAG)中為 N 且主計處行業代碼(CAREER_CODE)為"O-2"(公務人員)者改賦值為 Y。

■ 判斷小商號負責人客群

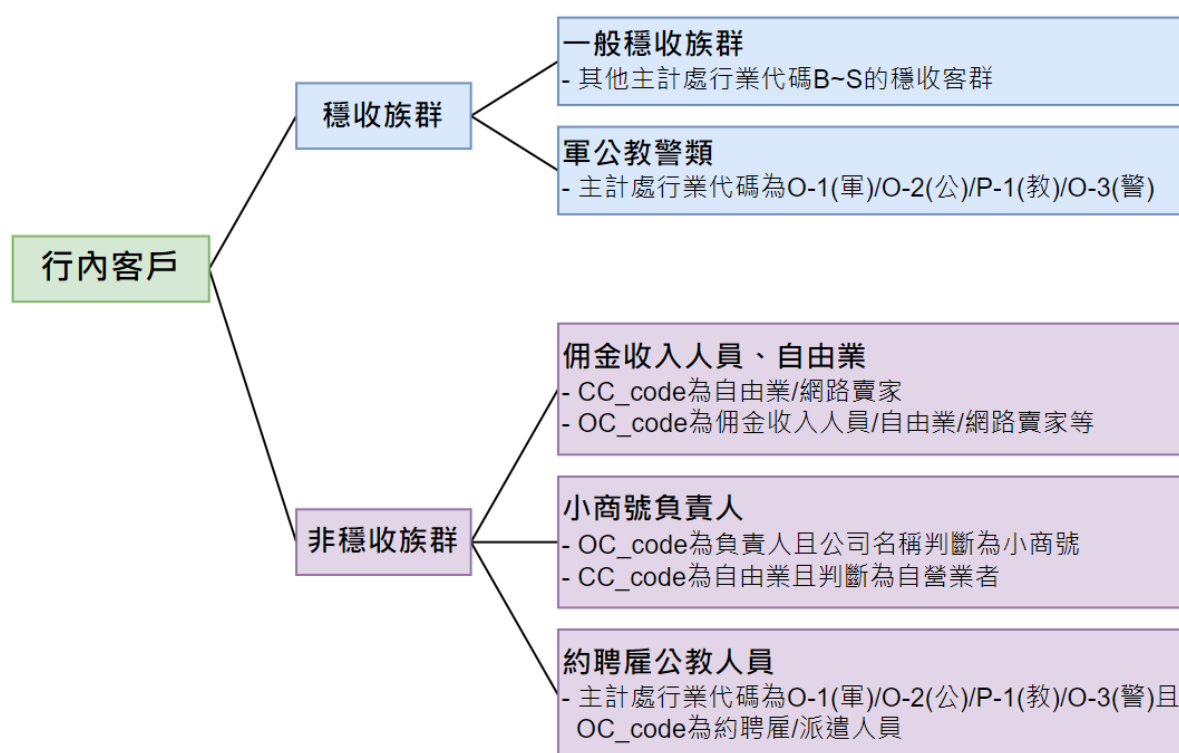
小商號負責人主要是從職稱別分類(OC_CODE_REVIEW)為 1, 27, 38 的樣本進行加工判斷，當其公司名稱(COMPANY)當中出現以下字串：「"公司", "事務所", "診所", "醫院", "藥局", "牙醫", "京城銀行", "昱嘉科技", "法國興業", "美祿科技", "直得科技"」時，則視為一般的負責人，負責人註記(CHAIRMAN_FLAG)賦值為"Firm"，其他則視為小商號負責人，賦值為"Store"。

另外有一部分小商號負責人則存在於自由業(CC_CODE_REVIEW = 34)的客戶中，處理邏輯如下：

1. 審查意見(credit_comment)中出現關鍵字：「"負責人", "自營", "店長"」但是公司名稱(COMPANY)不包含「"公司", "自營團購", "自營保姆", "自營包車司機"」的視為小商號負責人，賦值為"Store"

2. 公司名稱(COMPANY)包含「"自營飲料店","自營商","自營小吃"」的視為小商號負責人，賦值為"Store"

將不同客群也都貼上對應的標籤後，即可將行內客戶進行初步的分類。以薪資結構的觀點考量，軍公教警屬於穩收族群，其薪資多以本俸、年功俸以及各項專業加給與主管職務加給等組成，通常較為固定且不受景氣波動影響，相對於一般勞工之非經常性薪資可能會佔總薪資較高比例的結構有明顯差異，故單獨將軍公教警分出一類子模型。而以獎金或佣金收入為主的勞工、自營小商號的負責人、約聘僱的公教人員或代理代課教師則因收入形式較為不同且較不穩定，故而分類於非穩收族群的子模型。



針對一般穩收族群的部分，會透過分類演算法將薪資結構較接近的進行併組，有總體趨勢薪資的行業別共 17 類列示如下：

序號	主計處行業別代碼	主計處行業別名稱
1	B	礦業及土石採取業
2	C	製造業
3	D	電力及燃氣供應業
4	E	用水供應及污染整治業
5	F	營建工程業
6	G	批發及零售業
7	H	運輸及倉儲業

序號	主計處行業別代碼	主計處行業別名稱
8	I	住宿及餐飲業
9	J	出版影音及資通訊業
10	K	金融及保險業
11	L	不動產業
12	M	專業、科學及技術服務業
13	N	支援服務業
14	P	教育業(不含小學以上各級公私立學校等)
15	Q	醫療保健及社會工作服務業
16	R	藝術、娛樂及休閒服務業
17	S	其他服務業

考量本次建模資料蒐集期間起始年度為 2018 年，在行業別分類演算法之樣本區間選擇為利用 3.1.1 小節所產出之資料表 TABLE_SALARY_BASE，選取近五年從 2018 至 2022 年共 85 個樣本點的「平均月薪_全體(MONTH_INCOME_ALL)」欄位進行。行業別分類在決策樹演算法上為套用 rpart 套件的 rpart 函數²，因觀察值取值為連續型，將演算法所使用的方法參數(method)設定為 anova，其決策樹執行結果行業別依據薪資水準區分為 5 類；而 K-Means 演算法為套用基本統計套件中的 kmeans 函數並參考到決策樹的結果，將分類數參數設定為 5 類，最大疊代次數為預設值 10 進行行業別劃分。經比對兩個演算法結果可以看到多數行業別皆被分至相同類別中，僅有少數行業別因近一至兩年薪資有明顯提升而產生分類結果的差異。相關演算法結果請參考下表：

主計處行業別代碼	主計處行業別名稱	決策樹分類結果	K Means 分類結果
I	住宿及餐飲業	T10	K10
N	支援服務業	T10	K10
P	教育業(不含小學以上各級公私立學校等)	T10	K10
R	藝術、娛樂及休閒服務業	T10	K10
S	其他服務業	T10	K10
E	用水供應及污染整治業	T20	K20
F	營建工程業	T20	K20
G	批發及零售業	T30	K20、K30
L	不動產業	T30	K20、K30

² rpart 函數執行設定：

```
rpart(MONTH_INCOME_ALL ~ CAREER_CODE, data=Clustering_TB_5Y[, c("CAREER_CODE", "MONTH_INCOME_ALL")], method = "anova")
```


主計處行業別代碼	主計處行業別名稱	決策樹分類結果	K Means 分類結果
B	礦業及土石採取業	T30	K30
C	製造業	T30	K30
H	運輸及倉儲業	T30	K30
M	專業、科學及技術服務業	T40	K30、K40
Q	醫療保健及社會工作服務業	T40	K40
J	出版影音及資通訊業	T50	K40
D	電力及燃氣供應業	T50	K50
K	金融及保險業	T50	K50

除了兩個演算法皆依據薪資結構相近與否的分類結果，KPMG 亦考量到產業特性與實際行內建模資料集分佈，透過 3.4 節的資料合併結果檢視行內樣本各行業別的人數與平均月薪資統計，完成最終分類結果如下表格說明：

主計處行業別代碼	主計處行業別名稱	行內樣本數	行內資料之平均月薪資(元)	外部資料之五年平均月薪資(元)	最終分類結果說明
I	住宿及餐飲業	1,887	45,578	34,669	1. 依據決策樹與 K means 結果分類 2. 考量產業特性、薪資與資料樣本分佈，I、N、P、R、S 可歸於同一類型
N	支援服務業	1,346	50,839	37,618	
P	教育業(不含小學以上各級公私立學校等)	294	44,573	29,736	
R	藝術、娛樂及休閒服務業	422	52,083	39,105	
S	其他服務業	362	50,778	35,643	
E	用水供應及污染整治業	123	53,121	45,375	1. E、F 類為依據決策樹與 K means 結果分類 2. 經檢視 G 類批發及零售業，行內平均月薪資與 E、F 類相近，且在近年的外部薪資水準也較相近的情況下，將 G 類與 E、F 歸於同一類型
F	營建工程業	1,476	60,841	45,636	
G	批發及零售業	3,515	57,595	50,959	

主計處行業別代碼	主計處行業別名稱	行內樣本數	行內資料之平均月薪資(元)	外部資料之五年平均月薪資(元)	最終分類結果說明
L	不動產業	215	142,892	52,923	1. B、C、H 類依據決策樹與 K means 結果分類 2. 經檢視 L 類不動產類，在整體外部薪資水準較相近於 BCH 類的情況下包含近五年平均與最近一年月薪資，故將 L 類與 BCH 歸於同一類型
B	礦業及土石採取業	8	53,017	56,565	
C	製造業	12,990	88,922	54,925	
H	運輸及倉儲業	1,634	72,641	56,271	
M	專業、科學及技術服務業	1,015	137,726	63,137	M、Q 類屬較專業性質之客群，除依據決策樹與 K means 結果分類，實際檢視行內薪資結構相近且與其他客群有較大差異，故考量產業特性、薪資與資料樣本分佈，M、Q 類可歸與於同一類型
Q	醫療保健及社會工作服務業	2,836	171,229	64,956	
J	出版影音及資通訊業	965	85,038	73,332	1. 檢視 D 類電力及燃氣供應業行內薪資雖與外部薪資水準有落差，但仍屬薪資中上之客群，考量到外部薪資結構，建議仍依據決策樹與 K means 結果進行分類 2. 檢視 J 類資通訊業於行內外薪資統計皆屬較高薪資之客群。 3. 依上述說明，考量薪資結構與樣本分佈，可將 J、D、K 類歸於同一類型
D	電力及燃氣供應業	221	68,929	94,425	
K	金融及保險業	2,634	113,263	94,695	

最後透過單因子變異數分析檢視最終分類結果組與組之間的薪資結構是否有顯著差異，經變異數分析結果得出檢定統計量 $F = 230.7$ ，且 $p - value < 0.0001$ ，表示拒絕須無假設，不同組之間的薪資水準有顯著差異。

下表為最終行業別分類結果並納入軍公教警類、與非穩收族群子模型之相關描述性統計：

主計處行 業別代碼	主計處行業別 名稱	最終分類 結果	最終分類行 內樣本數	外部資料之最終 分類近五年平均 月薪資(元)	行內資料之最 終分類平均月 薪資(元)
I	住宿及餐飲業	一 般 服 務 業(C10)	4,311	35,354	48,226
N	支援服務業				
P	教育業(不含小 學以上各級公 私立學校等)				
R	藝術、娛樂及休 閒服務業				
S	其他服務業				
E	用水供應及污 染整治業	工 程 及 批 發 零 售 業 (C20)	5,114	47,323	58,424
F	營建工程業				
G	批發及零售業				
L	不動產業	技 術 勞 動 業(C30)	14,847	55,171	87,892
B	礦業及土石採 取業				
C	製造業				
H	運輸及倉儲業				
M	專業、科學及技 術服務業	專 業 服 務 業(C40)	3,851	64,046	162,398
Q	醫療保健及社 會工作服務業				
J	出版影音及資 通訊業	資 通 金 融 能 源 業 (C50)	3,820	87,484	103,568
D	電力及燃氣供 應業				
K	金融及保險業				
O-1	軍方機構	軍 公 教 警 類(C60)	2,870	NA	76,956
O-2	公務人員、政 府機構				
O-3	警務機構				
P-1	公立學校、私 立學校				
-	-	非 穩 收 族 群子模型	7,736	NA	75,290

3.3 建立總體趨勢薪資彙總表

在完成行業別分類後，最終分類結果將其區分為六類並統一名稱定義為子模型分類，後續依照子模型分類結果進行總體趨勢薪資計算，其運算過程為先建立子模型分類對照表並命名為 CAREER_GROUPING，其對照如下表：

主計處行業別代碼(CAREER_CODE)	子模型分類代碼(MODEL_CODE)
B	C30
C	C30
D	C50
E	C20
F	C20
G	C20
H	C30
I	C10
J	C50
K	C50
L	C30
M	C40
N	C10
P	C10
Q	C40
R	C10
S	C10
P-1	C60
O-1	C60
O-2	C60
O-3	C60

將上述子模型分類對照表 (CAREER_GROUPING) 之「子模型分類代碼 (MODEL_CODE)」，透過欄位「主計處行業別代碼(CAREER_CODE)」作為鍵值，與 3.1.1 小節所產出之資料表 TABLE_SALARY_BASE 的「行業代碼大類 (CAREER_CODE)」欄位進行比對後帶入該資料表中，後續再依據子模型分類代碼 (MODEL_CODE) 及資料年月 (DATA_YM) 兩個欄位歸戶後作加權統計³，計算各子模型

³ 加權統計為行業人數與月薪之加權

分類下之行業人數與加權後之平均月薪。以資料年月 202007 及子模型分類 C50 為例，先分別加總「加權月薪_全體(MONTH_INCOME_PPL_ALL)」⁴和「行業人數_全體(NUM_PPL_ALL)」⁵兩個欄位，再將加總後 MONTH_INCOME_PPL_ALL 除以加總後 NUM_PPL_ALL，即可得出子模型分類 C50 於 202007 時點之「平均月薪_全體(MONTH_INCOME_ALL_WA)」，而平均月薪_男(MONTH_INCOME_M_WA)及平均月薪_女(MONTH_INCOME_F_WA)之計算⁶亦同。經上述計算邏輯完成子模型分類 C10~C50 之總體趨勢子模型薪資彙總表並命名為 TABLE_SALARY_GROUP。相關資料表之彙整欄位及統計值說明如下：

序號	欄位名稱	欄位說明	欄位註解
1	MODEL_CODE	子模型分類代碼	
2	MODEL_NAME	子模型分類名稱	
3	DATA_YM	資料年月	
4	NUM_PPL_ALL	行業人數_全體	加總相同子模型分類之全體人數
5	NUM_PPL_M	行業人數_男	加總相同子模型分類之男性人數
6	NUM_PPL_F	行業人數_女	加總相同子模型分類之女性人數
7	MONTH_INCOME_ALL_WA	平均月薪_全體	加總相同子模型分類之加權月薪_全體/行業人數_全體
8	MONTH_INCOME_M_WA	平均月薪_男	加總相同子模型分類之加權月薪_男/行業人數_全體
9	MONTH_INCOME_F_WA	平均月薪_女	加總相同子模型分類之加權月薪_女/行業人數_全體

3.4 合併總體趨勢薪資彙總表與內部資料

本節為針對 3.1.2 & 3.3 節完成的子模型分類 C10~C50 及軍公教警 C60 之總體趨勢子模型薪資彙總表如何與個體差異子模型之建模資料集進行串接與合併作說明，相關執行步驟與內容詳述如下：

⁴ 3.1.1 小節所產出之資料表 TABLE_SALARY_BASE 欄位，即 p.17 表格中的序號 16，其公式為平均月薪_全體(MONTH_INCOME_ALL)乘以行業人數_全體(NUM_PPL_ALL)

⁵ 3.1.1 小節所產出之資料表 TABLE_SALARY_BASE 欄位，即 p.17 表格中的序號 7

⁶ 平均月薪_男(MONTH_INCOME_M_WA)之計算為加總後 MONTH_INCOME_PPL_M 除以加總後 NUM_PPL_M，而平均月薪_女(MONTH_INCOME_F_WA)之計算為加總後 MONTH_INCOME_PPL_F 除以加總後 NUM_PPL_F

■ 拆分穩收與非穩收族群資料集

首先，因外部薪資資料在穩收與非穩收族群有不同的資料處理方式，故在建模資料集初版(SourceTable)可先拆分穩收族群資料集與非穩收族群資料集。在穩收族群部分，拆分邏輯為篩選穩收註記(STABLE_FLAG)欄位為“Y”，且負責人註記(CHAIRMAN_FLAG)欄位為“Firm”或者空值者即為穩收族群，依上述邏輯完成資料拆分後輸出穩收族群資料集名為 ModelTable_Stable；而非穩收族群拆分邏輯為篩選穩收註記(STABLE_FLAG)欄位為“N”，或者負責人註記(CHAIRMAN_FLAG)欄位為“Store”者即為非穩收族群，依上述邏輯完成資料拆分後輸出非穩收族群資料集名為 ModelTable_Unstable。

■ 建立子模型分類於穩收族群資料集 ModelTable_Stable

本步驟將前述的穩收族群資料集(ModelTable_Stable)之主計處代碼(CAREER_CODE)欄位與 3.3 節建立的子模型分類對照表(CAREER_GROUPING)之 CAREER_CODE 欄位作為兩資料表的鍵值，比對後將子模型分類代碼(MODEL_CODE)欄位建立於穩收族群資料集 (ModelTable_Stable) 中，完成後輸出穩收族群暫存資料集名為 ModelTable_Stable_tmp。

■ 資料表合併-子模型分類之總體趨勢薪資彙總表，產出 ModelTable_Stable_v01 資料集

於資料表合併過程，須建立行業別資料匹配年月欄位(SALARY_Mapping_YM)，該欄位為透過資料對照年月(Mapping_YM)，將該欄位拆分為資料年度(YEAR)與資料月份(MONTH)並進行調整，其調整邏輯為考量子模型分類之總體趨勢薪資彙總表(TABLE_SALARY_GROUP)的外部資料來源更新時點會比實際更新的資料晚 2~3 個月，即 2021 年 7 月的統計薪資資料，實際公布時間可能為 2021 年 9 月或 10 月，並且該彙總表最終呈現是以年為單位並以 7 月為基準月所彙整之月薪資，因此在資料套用時點的原則建立為該年度 10 月份至隔年度 9 月份的資料可以取用該年度 7 月份所彙整的月薪資作該行業別下的總體趨勢薪資。

當資料對照年月(Mapping_YM)的月份(MONTH)大於等於 10 月時，則「行業別資料匹配年月(SALARY_Mapping_YM)」為資料年度(YEAR)與字串“07”合併而得，而當資料對照年月(Mapping_YM)的月份(MONTH)小於 10 月時，則「行業別資料匹配年月(SALARY_Mapping_YM)」為資料年度(YEAR)減去 1 後再與字串“07”合併而得。

綜合上述的資料處理與欄位衍生，將前述步驟之穩收族群暫存資料集 (ModelTable_Stable_tmp)透過子模型分類代碼(MODEL_CODE)篩選出 C10~C50 的樣本，後續基於篩選後的穩收族群資料集欄位子模型分類代碼(MODEL_CODE)與行業

別資料匹配年月(SALARY_Mapping_YM)·並以 3.3 節完成的子模型分類之總體趨勢薪資彙總表(TABLE_SALARY_GROUP)欄位子模型分類代碼(MODEL_CODE)與資料年月(DATA_YM)作為兩個資料表的合併主鍵值·比對後完成最終子模型分類之總體趨勢子模型薪資彙總表與建模資料集的合併並輸出建模資料集-1 名為 ModelTable_Stable_v01。

■ 資料表合併-軍公教警之總體趨勢薪資彙總表·產出 ModelTable_Stable_v02 資料集

軍公教警之總體趨勢薪資彙總表依據 3.1.2 節之說明可分為「公務人員薪資彙總表(DATA_PUBLIC)」、「教師薪資彙總表(DATA_TEACHER)」、「警察薪資彙總表(DATA_POLICE)」及「軍人薪資彙總表(DATA_MILITARY)」·針對該四張彙總表之合併須先利用前述步驟之穩收族群暫存資料集(ModelTable_Stable_tmp)分別建立「平均月薪_全體(MONTH_INCOME_ALL_WA)」、「平均月薪_男(MONTH_INCOME_M_WA)」及「平均月薪_女(MONTH_INCOME_F_WA)」欄位·後續篩選出子模型分類代碼(MODEL_CODE)為 C60 的樣本·透過該樣本中的資料年月(Mapping_YM)與年齡(AGE)欄位進行區間判斷·找出與符合資料區間的總體趨勢平均月薪資(含全體、男、女)分別套入上述建立的欄位中。因此於資料表合併過程·將依照下述邏輯及迴圈方式執行資料表合併：

若篩選後樣本集的主計處行業別代碼(CAREER_CODE)屬於 O-1·則利用樣本集的資料年月(Mapping_YM)與「軍人薪資彙總表(DATA_MILITARY)」的資料套用起始年月(BGN_YM)及資料套用結束年月(END_YM)進行區間判斷·而樣本集的年齡(AGE)則與「軍人薪資彙總表(DATA_MILITARY)」的年齡下界(AGE_BOT)及年齡上界(AGE_TOP)進行區間判斷。然而在軍人薪資彙總表(DATA_MILITARY)中·相同年齡區間有包含士兵、士官與軍官之薪資統計·因此從年齡區間判斷找出對應的三種不同位階之總體趨勢薪資欄位「平均月薪資_調整(MONTH_SALARY_ALL_WA)」須取平均值後作為最終之總體趨勢薪資·並將其套入「平均月薪_全體(MONTH_INCOME_ALL_WA)」欄位中。最後·因軍人薪資彙總表無相關性別薪資的統計·故僅串接全體平均月薪部分。

若篩選後樣本集的主計處行業別代碼(CAREER_CODE)屬於 O-2·則利用樣本集的資料年月(Mapping_YM)與「公務人員薪資彙總表(DATA_PUBLIC)」的資料套用起始年月(BGN_YM)及資料套用結束年月(END_YM)進行區間判斷·而樣本集的年齡(AGE)則與「公務人員薪資彙總表(DATA_PUBLIC)」的年齡下界(AGE_BOT)及年齡上界(AGE_TOP)進行區間判斷·找出對應的總體趨勢薪資欄位包含「全

體平均月薪資_調整(MONTH_SALARY_ALL_WA)」、「男平均月薪資_調整(MONTH_SALARY_M_WA)」及「女平均月薪資_調整(MONTH_SALARY_F_WA)」，並將欄位值分別套入「平均月薪_全體(MONTH_INCOME_ALL_WA)」、「平均月薪_男(MONTH_INCOME_M_WA)」及「平均月薪_女(MONTH_INCOME_F_WA)」欄位中。

若篩選後樣本集的主計處行業別代碼(CAREER_CODE)屬於 O-3，則利用樣本集的資料年月(Mapping_YM)與「警察薪資彙總表(DATA_POLICE)」的資料套用起始年月(BGN_YM)及資料套用結束年月(END_YM)進行區間判斷，而樣本集的年齡(AGE)則與「警察薪資彙總表(DATA_POLICE)」的年齡下界(AGE_BOT)及年齡上界(AGE_TOP)進行區間判斷，找出對應的總體趨勢薪資欄位包含「全體平均月薪資_調整(MONTH_SALARY_ALL_WA)」、「男平均月薪資_調整(MONTH_SALARY_M_WA)」及「女平均月薪資_調整(MONTH_SALARY_F_WA)」，並將欄位值分別套入「平均月薪_全體(MONTH_INCOME_ALL_WA)」、「平均月薪_男(MONTH_INCOME_M_WA)」及「平均月薪_女(MONTH_INCOME_F_WA)」欄位中。

若篩選後樣本集的主計處行業別代碼(CAREER_CODE)屬於 P-1，則利用樣本集的資料年月(Mapping_YM)與「教師薪資彙總表(DATA_TEACHER)」的資料套用起始年月(BGN_YM)及資料套用結束年月(END_YM)進行區間判斷，而樣本集的年齡(AGE)則與「教師薪資彙總表(DATA_TEACHER)」的年齡下界(AGE_BOT)及年齡上界(AGE_TOP)進行區間判斷，找出對應的總體趨勢薪資欄位包含「全體平均月薪資_調整(MONTH_SALARY_ALL_WA)」、「男平均月薪資_調整(MONTH_SALARY_M_WA)」及「女平均月薪資_調整(MONTH_SALARY_F_WA)」，並將欄位值分別套入「平均月薪_全體(MONTH_INCOME_ALL_WA)」、「平均月薪_男(MONTH_INCOME_M_WA)」及「平均月薪_女(MONTH_INCOME_F_WA)」欄位中。

根據上述迴圈判斷完之後，再針對代理/代課教師與約聘僱人員進行薪資水準的調整，調整方式如下：

1. 代理/代課教師

代理教師待遇接近正式老師，但寒暑假往往不支薪，且不受年薪考核，故原本預設 13.5 個月薪水應調整為 10 個月，而代課教師以目前的分類下無法判斷出來，故以代理教師的薪資結構計算，即當 CAREER_CODE 為“P-1”(教師)且 OC_CODE_REVIEW 為 6 或 55 時，將 MONTH_INCOME_ALL_WA 除以 13.5 再乘以 10

2. 約聘僱人員

約僱人員最高支給 5 等 280 薪點，而約聘人員則是最低從 6 等一階 280 薪點起算，遂以 280 薪點換算其薪資水準，由於約聘僱人員比照公務人員同享 1.5 個月薪資，故將 280 薪點乘以每個薪點折合金額再乘以 13.5 除以 12 換算為月收入，2022/1/31 以前適用每薪點 124.7 元，故 APPROVE_DATE 在 2022/1/31 以前的樣本認定薪資水準為 39,280.5，2022/2/1 以後薪點折合金額調升 5 元，故 APPROVE_DATE 在 2022/2/1 以後的樣本認定薪資水準為 40,855.5

根據上述邏輯及程序完成最終軍公教警之總體趨勢子模型薪資彙總表與建模資料集的合併並輸出建模資料集-2 名為 ModelTable_Stable_v02。

■ 資料表合併-產出 ModelTable_Unstable 資料集

首先建立 TABLE_UnstableIncome_BASE 資料表，目的是將每一年各行業的總體平均薪資計算出來後存入表中，具體說明如下：

以迴圈遍歷資料年月(DATA_YM)201107~202207，在每個資料年月中

1. 加總 TABLE_SALARY_BASE 表的全體加權月薪(MONTH_INCOME_PPL_ALL)，再除以全體人數(NUM_PPL_ALL)，得到各行業的總體平均薪資當作當年度的總體平均月薪(MONTH_INCOME_ALL_WA)
2. 加總 TABLE_SALARY_BASE 表的女性加權月薪(MONTH_INCOME_PPL_F)，再除以全體女性人數(NUM_PPL_F)，得到各行業的女性總體平均薪資當作當年度的女性總體平均月薪(MONTH_INCOME_F_WA)
3. 加總 TABLE_SALARY_BASE 表的男性加權月薪(MONTH_INCOME_PPL_M)，再除以全體男性人數(NUM_PPL_M)，得到各行業的男性總體平均薪資當作當年度的男性總體平均月薪(MONTH_INCOME_M_WA)

然後，當 ModelTable_Unstable 的資料對照年月(Mapping_YM)的月份(MONTH)大於等於 10 月時，則「行業別資料匹配年月(SALARY_Mapping_YM)」為資料年度(YEAR)與字串"07"合併而得，而當資料對照年月(Mapping_YM)的月份(MONTH)小於 10 月時，則「行業別資料匹配年月(SALARY_Mapping_YM)」為資料年度(YEAR)減去 1 後再與字串"07"合併。

接下來就可以將 ModelTable_Unstable 與 TABLE_UnstableIncome_BASE 進行合併，即當 SALARY_Mapping_YM 與 DATA_YM 相同時進行 JOIN，並將資料集合併為 ModelTable_Unstable。

另外，ModelTable_Unstable 也跟 TABLE_SALARY_BASE 進行 JOIN，目的是為了得到每個主計處行業代碼下的薪資水準，故將 CAREER_CODE 相同且 SALARY_Mapping_YM 與 DATA_YM 相同的資料進行 JOIN，並將 MONTH_INCOME_ALL 的欄位改名為 MONTH_INCOME_CAREERCODE。

3.5 職類別薪資資料處理

針對外部資料處理部分，除了主計處薪情平臺與軍公教警人事管轄機關之公開資訊外，關於勞動部職類別薪資調查動態查詢資料亦於本章節進行說明。

為了利用勞動部職類別薪資調查動態查詢的外部薪資統計資料，首先需要建立職稱別分類(OC_CODE_REVIEW)與勞動部職稱代碼(JOB_CODE)的對照關係，但考量到行內的分類規則下較難以直接與勞動部職稱代碼進行匹配，為避免分類不當而導致錯誤套用，針對行內外職稱別分類的對照需進行妥適處理，故當部份情況下無法單獨以職稱別分類(OC_CODE_REVIEW)決定對應的勞動部職稱代碼(JOB_CODE)，仍需搭配公司名稱(COMPANY)或審查意見(credit_comment)中的關鍵字以判斷合適的分類，故 KPMG 綜合考量現行分類下的樣本結果，將初步處理規則彙整於附錄 5.1 行業與職稱別對照表「職稱 OC_Code」中的處理說明欄位與備註欄位，處理原則簡要說明如下：

依據歸類後之職稱代碼(JOB_CODE)名稱找出能與勞動部職類別名稱大致相符之職稱。舉例如：歸類後之職稱代碼 07_1，名稱為醫師，於勞動部職類別列表中同樣能找到名稱為醫師，代碼為 221090 的外部職稱。

若無法找到大致相符的職稱，則依照勞動部職稱說明⁷搭配職類別列表匹配出最適當之內外部職稱代碼對應。舉例如：歸類後之職稱代碼 06_10，名稱為技術員/作業員，檢視勞動部職類別列表中，技術員/作業員較符合外部職稱代碼 30 與 70 開頭項下的職稱，因此在建立歸類後之職稱代碼 06_10 與外部職稱對應上會套用代碼 300000，名稱為技術員及助理專業人員，以及代碼 700000，名稱為技藝、機械設備操作及組裝人員之職稱。後續依據對照結果完成內外部職稱與職稱倍率彙總表之計算。

⁷ 詳見附錄 5.3 勞動部職業名稱及工作說明

■ 內外部職稱與職稱倍率彙總表，命名為 TABLE_JOB

綜合上述，先以彙整好之職類別薪資調查動態查詢薪資彙整表⁸(以下稱動態薪資彙整表)中，資料年月(DATA_YM)調整成西元日期，西元年加月呈現六碼字元，命名為資料日期(DATA_YM)，並拆分年(YEAR)與月(MONTH)欄位，後續以內外部職稱代碼對照表為基礎，依序將歸類後之職稱代碼(JOB_CODE)所對應到之「外部職稱對應代碼」(若外部職稱對應代碼為 NA，則以 888800 代入)為鍵值，比對回動態薪資彙整表中欄位「JOB_TITLE_CODE」，並考量不同的資料日期(DATA_YM)下，找出各年度行業代碼大類(CAREER_CODE)之行業人數(NUM_PPL_ALL)及總薪資(月) (INCOME_ALL)兩欄。上述計算的程序實際操作如下示例：

以 CAREER_CODE=B 且 DATA_YM=201107 為例

- 當歸類後之職稱代碼為 04_1 時，對應到的外部職稱對應代碼為 888800，因此與動態薪資彙整表進行代碼匹配，會得到對應的行業人數與總薪資(月)，並作為歸類後之職稱代碼 04_1 的行業人數與總薪資(月)。

CAREER_CODE	CAREER_CH	DATA_YM	JOB_TITLE_CODE	JOB_TITLE_CH	NUM_PPL_ALL	INCOME_ALL
B	礦業及土石採取業	201107	888800	總計	3,978	44,078

- 當歸類後之職稱代碼為 06_10 時，對應到的外部職稱對應代碼為 300000 和 700000，因匹配超過 2 個以上之外部職稱，應計算其加總後之行業人數以及各外部職稱加權後之總薪資，作為歸類後之職稱代碼 06_10 所對應之行業人數與總薪資(月)

CAREER_CODE	CAREER_CH	DATA_YM	JOB_TITLE_CODE	JOB_TITLE_CH	NUM_PPL_ALL	INCOME_ALL
B	礦業及土石採取業	201107	300000	技術員及助理專業人員	194	59,087
B	礦業及土石採取業	201107	700000	技藝、機械設備操作及組裝人員	2,294	42,009
					加總後之NUM_PPL_ALL	加權後之INCOME_ALL
					2,488	43,341

將歸類後之職稱代碼所匹配到的總薪資(月) (INCOME_ALL)去除以該行業別代碼大類和資料日期下之 JOB_TITLE_CODE 為 888800 總計薪資，求得出職稱倍率_初始 (BENCHMARK_RATE)。以前述示例繼續說明，當歸類後之職稱代碼為 04_1 時，總薪資(月)為 44,078 元，則除以該資料日期下之 888800 總計薪資，則職稱倍率_初始為 $1(=44078/44078)$ ；當歸類後之職稱代碼為 06_10 時，總薪資(月)為 43,341 元，則除以

⁸ 詳見「王道銀行所得代理模型專案 模型開發報告書_資料處理篇」之 2.2.3 章節 & 「王道銀行所得代理模型專案 模型開發報告書_資料處理篇」附錄 5.4 EXCEL 檔之 MOL Source 頁籤

該資料日期下之 888800 總計薪資，職稱倍率_初始為 0.9833(=43341/44078)。此外，當實際職稱倍率_初始(BENCHMARK_RATE)計算為 0，表示在勞動部的資料中該產業別下無此職稱之薪資統計，故資料往後執行會採以外部職稱對應代碼為 888800 之總計薪資的職稱倍率(其值為 1)水準進行取代。

最終透過上述計算輸出各行業別於各資料日期下之內外部職稱與職稱倍率彙總表，並命名為 TABLE_JOB。

內外部職稱與職稱倍率彙總表欄位中英文名稱彙整如下：

序號	欄位名稱	欄位中文名稱
1	CAREER_CODE	行業別代碼大類
2	CAREER_CH	行業別中文名稱
3	JOB_TITLE_CODE_BANK	歸類後之職稱代碼
4	DATA_YM	資料日期
5	YEAR	資料年度
6	MONTH	資料月份
7	NUM_PPL_ALL	行業人數
8	INCOME_ALL	總薪資(月)
9	BENCHMARK_RATE	職稱倍率_初始

4. 個體差異子模型開發

依據 2.2 節之個體差異子模型開發方法說明，個體差異子模型開發包含模型變數清單、單變量分析、多變量分析以及模型驗證。模型資料集和變數清單部分已於模型開發報告書_資料處理篇中介紹相關內部資料之串接彙整與變數處理的流程，最終產出 **AnalyticalTable** 作為後續個體差異子模型開發的資料集。同時本章節除涵蓋模型開發內容外，亦包含模型開發前之樣本篩選、資料處理與樣本抽樣等處理，以下針對各步驟進行說明：

4.1 樣本排除條件

依據個體差異子模型的建置方法與目標篩選出合適的樣本範圍，並排除可能出現的極端值與無效樣本等。下表為個體差異子模型的樣本排除條件(即非信貸客群)：

排除條件	排除原因	原始 資料筆數	排除 資料筆數	剩餘 資料筆數
農林漁牧業 (CAREER_CODE = 'A') 家管/學生/義務役/退休/專業 投資人/房東/無業	主計處針對該行業類別 無相關薪資統計，無法 進行外部總體趨勢薪資 彙整，故排除於模型適 用範圍	43,334	200	43,134
根據行內授信客群 H 定義之 CC_CODE_REVIEW 與 OC_CODE_REVIEW	客群 H 屬高風險的負面 表列職業，根據貴行規 範不予承作，故排除於 模型適用範圍			

關於上述排除條件的具體定義如下表：

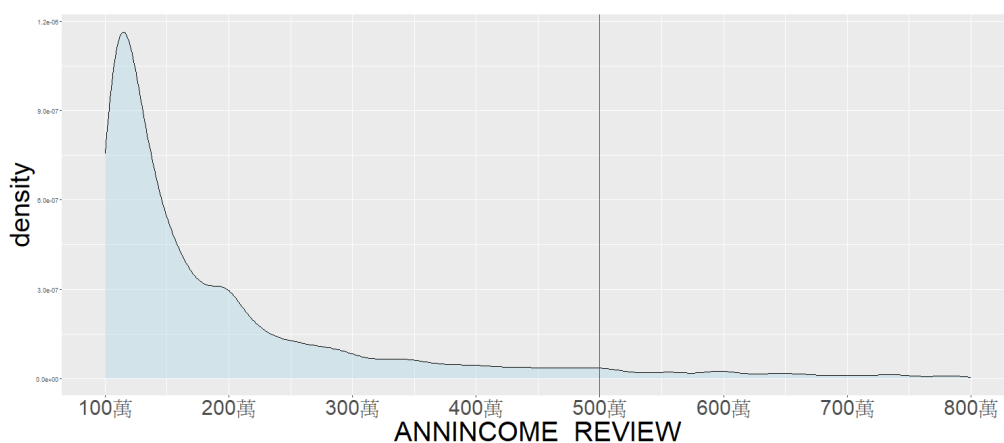
CC_CODE_REVIEW	OC_CODE_REVIEW	其他條件
31 合法登記娛樂場所、休閒 場所、其他服務業	26 佣金收入專業人員	
34 自由業		credit_comment 欄位中包含 ("退休", "投資", "房東")字串
39 軍火/武器之製造買賣業		
43 遊藝、博弈業		
44 當舖、民間融資、民間匯		

CC_CODE_REVIEW	OC_CODE_REVIEW	其他條件
款業		
49 軍火/武器之製造買賣業		
67 虛擬資產業		
68 特定娛樂業(三溫暖、舞廳酒店等聲色行業)		
69 退休		
70 家管		
71 無業		
72 學生		
	14 義務役少尉軍官	
	15 義務役士官、士兵、學生	
	30 退休、家管	
	31 退休	
	35 學生	
	67 家管	
	68 退休	

由於建模樣本中存在極端值時，可能會導致模型無法合理的描繪應變數與自變數的真實關係，以致於模型的預測結果容易產生較大的偏離，因此須設定合理的極端值排除條件。首先，根據民國 112 年 1 月 1 日勞動部最新發布的每月最低基本工資為 26,400 元，換算為年薪約為 30 萬，故以核實年收入 30 萬做為合理的下界排除門檻。

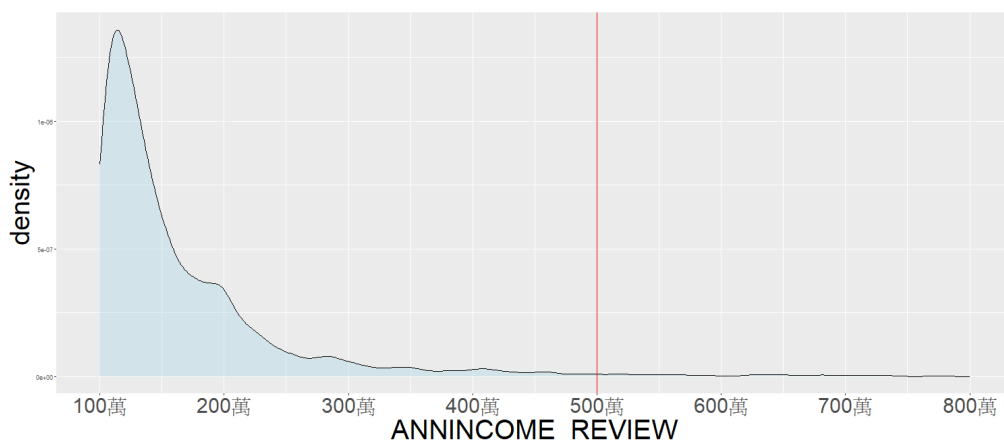
其中，穩收族群(C10~C60)約有 3.25%，共 1,130 筆樣本之核實年收入低於 30 萬，而非穩收族群則約有 4.97%，共 519 筆樣本之核實年收入低於 30 萬。這些樣本將會排除於建模樣本之外。

至於核實年收入的上界排除門檻則可以根據樣本分布來看，下圖是穩收族群的核實年收入右尾分布圖：



圖中顯示核實年收入大約至 500 萬以後分布開始趨緩收斂，僅約 719 筆樣本(約 2.06%)核實年收入高於 500 萬。

下圖則是非穩收族群的核實年收入右尾分布圖：



圖中顯示核實年收入也大約為 500 萬以後分布開始趨緩收斂，僅約 66 筆樣本(約 0.32%)核實年收入高於 500 萬。

綜上所述，將以核實年收入介於 30~500 萬的樣本設定為個體差異子模型合適的建模樣本範圍，最終各子模型樣本分布統計彙整如下：

子模型分類結果	子模型分類名稱	建模樣本拆分筆數
C10	一般服務業	4,007
C20	工程及批發零售業	4,717
C30	技術勞動業	14,363
C40	專業服務業	3,290
C50	資通金融能源業	3,724
C60	軍公教警類	2,870
非穩收族群		7,719

子模型分類結果	子模型分類名稱	建模樣本拆分筆數
	總計	40,690

■ 模型適用客群

個體差異子模型可適用的基本客群亦應與上述條件所篩選的樣本範圍趨於一致，故定義適用客群可歸納說明為「行內行業別代碼能跟主計處行業別 B、C、D、E、F、G、H、I、J、K、L、M、N、P、P-1、Q、R、S、O-1、O-2、O-3 類進行對應，不屬於客群 H，亦非農林漁牧業/家管/學生/義務役/退休/專業投資人/房東/無業者」。

4.2 應變數定義與分析

4.2.1 應變數定義

建置個體差異子模型之目的在於透過統計分析方法，尋找出各子模型分類下利用職稱、年齡、性別與聯徵資料等相關變數對應該客群收入距離之線性關係式，以期能利用多元線性迴歸模型進行收入距離的預測。

考量到自變數與收入距離之間是在不同的連續變數區間或是不同的類別等級下具有相異的增減幅，因此透過自然對數轉換的方式處理收入距離，作為個體差異子模型的應變數，以捕捉收入距離相對於自變數之間可能隱含的比率關係。

收入距離公式如下：

$$\begin{aligned}
 \text{收入距離[Distance_Income_Log]} \\
 &= \log(\text{月收入[Monthly_Income]}) \\
 &\quad - \log(\text{子模型分類下總體趨勢薪資[MONTH_INCOME_ALL_WA]})
 \end{aligned}$$

4.3 長清單定義

職位價值模型主要採用「王道銀行所得代理模型專案 模型資料處理與初步分析報告書」中附錄 5.6 變數長清單。各類型長清單變數摘要如下：

變數類型	變數說明	變數個數
客戶基本資料	彙整信貸系統與房貸系統之客戶基本資料包含年齡、學歷、行業別及居住地區域等。	21
聯徵查詢資料	本次聯徵查詢資料包含 B 類及 K 類。B 類主要為授信產品之繳款行為及信用額度等相關資訊，而 K 類主要為信用卡之持卡狀況、一般消費及信用卡額度等相關資訊。	44
行內往來資料	存款往來明細資訊，如平均每月出入帳金額、交易次數、近數個月內出入帳金額高於幾萬的次數。	60

4.4 抽樣設計與結果

KPMG 在樣本抽樣設計中，為確保每一群不同收入級距的客戶都具有足夠代表性的樣本，其作法為將核實年收入(ANNINCOME_REVIEW)切分成多個區間進行分層抽樣，於抽樣過程中將各區間 80%樣本作為樣本內的訓練集來建模，並保留 20%樣本作為樣本外的測試集來驗證。其中個體差異子模型將年收入分層如下：「30~40(含)萬」、「40~50(含)萬」、「50~75(含)萬」、「75~100(含)萬」、「100~150(含)萬」、「150~200(含)萬」、「200~250(含)萬」、「250~300(含)萬」、「300~400(含)萬」、「400~500(含)萬」，共 10 組。

下表為利用建模樣本⁹經過子模型拆分所取得之各子模型的建模樣本集，後續再將各子模型的建模樣本集進行分層抽樣的結果：

子模型分類	建模樣本拆分筆數	訓練集樣本數	測試集樣本數
C10	4,007	3,205	802
C20	4,717	3,773	944
C30	14,363	11,490	2,873
C40	3,290	2,632	658
C50	3,724	2,979	745

⁹ 建模樣本定義為完成 4.1 節樣本排除條件之排除序號 1~9 步驟後產出之資料集

子模型分類	建模樣本拆分筆數	訓練集樣本數	測試集樣本數
C60	2,870	2,296	574
非穩收族群	7,719	6,174	1,545
總計	40,690	32,549	8,141

4.5 單變量分析結果

單變量分析之目的是為了從全部變數中篩選出較具檢定力與代表性的變數，再進行後續的多變量分析與模型開發。具體的分析資料集與步驟說明如下：

1. 檢視各子模型樣本敘述統計值，包含缺失值與零值的比例、最大值與最小值、唯一值筆數、眾數、四分位數、第 1 與第 99 百分位數、平均值、標準差、離群值、正值與負值筆數等，確定樣本篩選範圍與排除條件。
2. 計算相關係數、MAPE，先後以 MAPE 與相關係數進行排序，並考量缺失值或零值比例在合理情況下，優先挑選 MAPE 較低、相關係數較高的變數作為短清單變數。

相關敘述統計與單變量分析結果請參考附錄 5.5 敘述統計與單變量分析。而經單變量分析後，變數個數彙整如下：

子模型分類	長清單變數個數	短清單變數個數
C10	135 個	23 個
C20	135 個	23 個
C30	135 個	23 個
C40	135 個	22 個
C50	135 個	23 個
C60	132 個	23 個
非穩收族群	135 個	27 個

4.6 一般服務業(C10)子模型多變量分析與模型驗證結果

針對子模型分類 C10(一般服務業)所應用到的資料集說明請參考 4.4 節，而應用到的資料集及結果輸出名稱列示如下：

資料集說明	資料集名稱
訓練集	TrainData_C10_Output.RData
測試集	TestData_C10_Output.RData
建模樣本	ModelData_C10_Output.RData

4.6.1 模型設定

收入距離 = $\alpha + \beta_1$ 職稱倍率 + β_2 推估總年資註記 + β_3 工作地年薪平均數 + β_4 性別倍率
+ β_5 無擔保貸款總額度 + β_6 近 12 個月平均每發卡銀行永久額度
+ β_7 有擔保貸款高額度註記 + β_8 近 6 個月高消費註記 + ε

4.6.2 模型結果

以 3,205 筆訓練集資料配適多元線性迴歸，模型係數四捨五入至小數第 8 位，C10 子模型結果如下：

變數名稱	模型係數	T 值	P 值	VIF 值
截距項	-1.30201226	-15.78	<0.01	-
職稱倍率 Jobtitle_Rate	0.15976769	19.76	<0.01	1.09
推估總年資註記 SENIORITY_Mark_TOT	0.10078284	7.95	<0.01	1.10
工作地年薪平均數 WORK_AVG	0.00298013	7.17	<0.01	1.02
性別倍率 Gender_Rate	0.84019041	10.95	<0.01	1.02
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00208116	21.24	<0.01	1.18
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_10K_n1	0.01483032	14.47	<0.01	1.35

變數名稱	模型係數	T 值	P 值	VIF 值
有擔保貸款高額度註記 HighLimit_Secur_Mark_C10	0.15282179	7.60	<0.01	1.10
近 6 個月高消費註記 HighSpending_6M_Mark_C10	0.15133077	9.36	<0.01	1.20

C10 子模型的 R-squared 值為 0.4971，Adjusted R-squared 值為 0.4958，所有投入變數之模型係數均顯著異於 0。VIF 值均小於 10，變數之間不存在共線性問題。

在業務的應用上，會希望能夠檢視各變數的重要程度，以了解其貢獻性。因此 KPMG 將上述模型進行標準化迴歸，其迴歸係數便能夠揭露其重要程度。結果與變數對應的權重列表如下：

變數名稱	標準化迴歸係數	變數權重
職稱倍率 Jobtitle_Rate	0.2591	19.6%
推估總年資註記 SENIORITY_Mark_TOT	0.1048	7.9%
工作地年薪平均數 WORK_AVG	0.0908	6.9%
性別倍率 Gender_Rate	0.1384	10.5%
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.2895	21.9%
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_10K_n1	0.2109	15.9%
有擔保貸款高額度註記 HighLimit_Secur_Mark_C10	0.1001	7.6%
近 6 個月高消費註記 HighSpending_6M_Mark_C10	0.1286	9.7%

根據 C10 子模型的模型結果，投入對應的變數即可預測收入距離，後續透過 4.2 節收入距離公式可推導出預測月收入¹⁰，結果詳見資料集中的「yhat」欄位。

¹⁰ 依模型估計得到的預測收入距離(\hat{y}_d)會以 $e^{\hat{y}_d + \log(\text{子模型分類下之總體趨勢薪資}[\text{MONTH_INCOME_ALL_WA}])}$ 轉換為預測月收入

4.6.3 模型效度

下表為子模型分類 C10 的誤差百分比¹¹在各區間的樣本佔比以及 MAPE¹²統計結果：

誤差百分比	訓練集	測試集	新增測試集 ¹³	全樣本
-100%~-50%	1.87%	2.24%	2.33%	1.95%
-50%~-20%	16.51%	14.84%	11.01%	16.17%
-20%~-10%	13.07%	12.34%	9.24%	12.93%
-10%~0%	15.66%	14.71%	11.01%	15.47%
0%~10%	14.57%	14.71%	10.91%	14.60%
10%~20%	12.61%	13.47%	10.07%	12.78%
20%~40%	17.25%	18.45%	19.22%	17.49%
40%~50%	3.68%	4.11%	8.86%	3.77%
50%~75%	3.21%	3.62%	13.43%	3.29%
75%~100%	1.00%	0.50%	2.15%	0.90%
>100%	0.56%	1.00%	1.77%	0.65%
樣本整體 MAPE	21.61%	22.55%	31.43%	21.80%
樣本數	3,205	802	1,072	4,007

誤差範圍	訓練集	測試集	新增測試集	全樣本
<-10 萬	0.56%	0.87%	1.31%	0.62%
-10 萬~-5 萬	1.68%	1.62%	1.12%	1.67%
-5 萬~-3 萬	2.81%	3.12%	2.33%	2.87%
-3 萬~-2 萬	4.06%	3.12%	2.33%	3.87%
-2 萬~-1 萬	10.73%	9.98%	7.46%	10.58%
-1 萬~0 萬	27.27%	25.44%	19.03%	26.90%
0 萬~1 萬	38.53%	40.15%	36.57%	38.86%
1 萬~2 萬	11.01%	12.34%	25.75%	11.28%
2 萬~3 萬	1.78%	1.75%	2.43%	1.77%
3 萬~5 萬	1.09%	1.00%	1.12%	1.07%
5 萬~10 萬	0.41%	0.50%	0.47%	0.42%

¹¹ 誤差百分比 = (預測月收入-核實月收入) / 核實月收入

¹² 將「yhat」欄位的預測月收入，與「Monthly_Income」欄位的核實月收入套入 MAPE 的計算公式(詳見 2.2.4 節之模型效度指標)，即可計算出預測月收入與核實月收入間的 MAPE

¹³ 新增測試集為 2023/12/5 依貴行提出需求新增之測試集，擷取樣本範圍是訓練集以外樣本，並排除(1)非信貸客群 (2)無擔總額度> 500 萬之非適用客群 (3)策略預估大於年收>200 萬之非適用客群 (4)貴行授信政策下不予承做之核實年薪 24 萬以下客群

誤差範圍	訓練集	測試集	新增測試集	全樣本
>10 萬	0.06%	0.12%	0.09%	0.07%
平均絕對誤差(元)	11,333.08	11,816.75	15,871.44	11,429.88

由上表結果顯示，C10 子模型大約有 94% 的樣本估計誤差介於 $\pm 50\%$ 之間，而平均絕對誤差¹⁴ 約為 11,333 元。每筆樣本的誤差百分比之詳細結果詳見資料集中的「PerErr」欄位，誤差值詳見「Err」欄位。

■ 各收入區間¹⁵

下表是將全樣本分隔成多個預估年收入區間，並計算在各區間樣本下不同誤差範圍下的樣本占比，以及各區間的樣本數、平均絕對誤差(MAE)、平均絕對誤差百分比(MAPE)。後續的應用策略框架報告書將會再說明 C10 子模型與分量迴歸模型的搭配使用。

誤差範圍 (萬元)	預估年收入區間(萬元)												總計
	0-30	30-40	40-50	50-75	75-100	100-150	150-200	200-250	250-300	300-400	400-500	> 500	
<-10	0.0%	0.0%	0.0%	0.2%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.6%
-10~-5	0.0%	0.0%	0.2%	0.5%	0.5%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	1.7%
-5~-3	0.0%	0.0%	0.8%	1.3%	0.3%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	2.9%
-3~-2	0.0%	0.0%	1.1%	2.2%	0.4%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3.9%
-2~-1	0.0%	0.7%	4.5%	4.3%	0.7%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	10.6%
-1~0	0.0%	3.1%	14.5%	8.2%	0.8%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	26.9%
0~1	0.0%	4.7%	22.5%	10.7%	0.7%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	38.9%
1~2	0.0%	0.0%	3.4%	6.7%	0.8%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	11.3%
2~3	0.0%	0.0%	0.0%	1.0%	0.4%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.8%
3~5	0.0%	0.0%	0.0%	0.1%	0.4%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.1%
5~10	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.4%
>10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%
樣本數	0	344	1,890	1,407	210	110	32	6	5	1	2	0	4,007
MAE	-	5,044	7,390	12,459	26,747	40,488	42,054	58,921	94,120	79,965	124,072	-	11,430
MAPE	-	14.8%	18.4%	24.1%	33.4%	40.1%	47.5%	28.7%	36.1%	45.7%	46.4%	-	21.8%

¹⁴ 平均絕對誤差(MAE)即為每一筆樣本預測月收入與核實月收入相減後取絕對值再平均

¹⁵ 各收入區間是以年收入(ANNINCOME_REVIEW)欄位進行判斷，根據年收入高低劃分成 12 個不同所得級距的區間。

4.6.4 模型變數相關說明

■ 職稱倍率

變數名稱為 Jobtitle_Rate，其變數來源由建模資料集與 3.5 節職類別薪資資料處理產出之內外部職稱與職稱倍率彙總表即 TABLE_JOB，利用下述欄位與邏輯進行資料對應與串接而得：

- 職稱資料匹配年月欄位(ADJ_Mapping_YM)，該欄位為透過資料對照年月(Mapping_YM)，將該欄位拆分為資料年度(YEAR)與資料月份(MONTH)並進行調整後合併，其調整邏輯為考量內外部職稱與職稱倍率彙總表的外部資料來源更新時點為每年度的 5 月底至 6 月初，而公布的資料為去年度的 7 月份，舉例如 2022 年 5 月底發布的外部資料為 2021 年 7 月各行業別下各職稱的薪資，因此 2022 年 7 月至 2023 年 6 月的樣本可套用 2021 年 7 月的職稱薪資統計資料，故在建立內外部職稱與職稱倍率彙總表的「職稱資料匹配年月欄位(ADJ_Mapping_YM)」上，當資料對照年月(Mapping_YM)的月份(MONTH)大於等於 7 月時，職稱資料匹配年月(ADJ_Mapping_YM)為資料年度(YEAR)減去 1 後再與字串"07"合併而得，反之當資料對照年月(Mapping_YM)的月份(MONTH)小於 7 月時，職稱資料匹配年月(ADJ_Mapping_YM)為資料年度(YEAR)減去 2 後再與字串"07"合併而得，最後利用上述產出之職稱資料匹配年月(ADJ_Mapping_YM)符合資料年月(DATA_YM)的邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同年度的資料。
- 主計處行業別代碼(CAREER_CODE)，可由 3.2 節之說明以重新歸類後之行業別與邏輯取得，利用該欄位符合行業別代碼大類(CAREER_CODE)邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同行業別代碼的資料。
- 職稱(JOB_CODE)，可由 3.5 節之說明以重新歸類後之職稱別與邏輯取得，利用該欄位符合行內職稱代碼(JOB_TITLE_CODE_BANK)邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同職稱代碼的資料。

針對上述欄位與邏輯利用迴圈方式同時進行判斷，找出並串接每一筆資料集樣本於內外部職稱與職稱倍率彙總表(TABLE_JOB)中相應的職稱倍率_初始(BENCHMARK_RATE)欄位，最後將該欄位命名為 Jobtitle_Rate。

■ 推估總年資註記

變數名稱為 SENIORITY_Mark_TOT，變數來源為建模資料集所彙整的「年齡(AGE)」與「教育程度(EDUCATION)」欄位，該變數相關處理邏輯參考如下：

- 若教育程度(EDUCATION)屬於 0(研究所以上)者，則將年齡(AGE)減去 24 作為推估總年資(SENIORITY_YEAR_TOT)變數值
- 若教育程度(EDUCATION)屬於 1(大學/大專)者，則將年齡(AGE)減去 22 作為推估總年資(SENIORITY_YEAR_TOT)變數值
- 若教育程度(EDUCATION)屬於 2(高中)者，則將年齡(AGE)減去 18 作為推估總年資(SENIORITY_YEAR_TOT)變數值
- 若教育程度(EDUCATION)屬於 3(國中以下)者，則將年齡(AGE)減去 15 作為推估總年資(SENIORITY_YEAR_TOT)變數值

實際以上述邏輯方式推估，推估總年資會有 1~2 年誤差，舉例如教育程度(EDUCATION)屬於 0(研究所以上)，但年齡(AGE)為 22 者，如遇該情況(於 C10 子模型共 18 筆)則以 0 值做取代。後續檢視並參考該變數百分位點之數值並反覆測試模型找到最佳之高年資門檻，當變數值大於等於門檻值則賦值為 1，小於門檻值則賦值為 0，而最終額度門檻訂為 20 年。

■ 工作地年薪平均數

變數名稱為 WORK_AVG，變數來源為資料集所彙整的「公司郵遞區號(POST_CODE_COMPANY)」欄位。該變數相關處理是以行政院主計總處於 109 發布之「各縣市工業及服務業企業僱用員工薪資統計說明」¹⁶中表 1，以及 110 年所發布之「工業及服務業受僱員工全年總薪資中位數及分布統計結果」¹⁷中表 2-1「各縣市工業及服務業企業僱用員工全年總薪資」統計之年薪資平均數(單位：萬元)作為薪資水準進行套用，資料期間共包含了 107~110 年度，KPMG 也將其數據彙整於「特徵變數轉換檔」¹⁸的地理區域變數總表頁籤中。

¹⁶ 「各縣市工業及服務業企業僱用員工薪資統計說明」請參考附錄 5.4 之 PDF 檔

¹⁷ 「工業及服務業受僱員工全年總薪資中位數及分布統計結果」請參考附錄 5.5 之 PDF 檔

¹⁸ 「特徵變數轉換檔」的地理區域變數總表請參考附錄 5.6 之 EXCEL 檔

考量到該份外部資料更新時點為隔年 12 月底，例如：2021 年底發布的外部資料為 2020 年度的薪資統計，因此 2022 年 1 月至 2022 年 12 月的樣本可套用最新發布之 2020 年度的薪資統計，故以下列規則衍生變數：

- APPROVE_DATE 在 2018/01/31~2020/12/31，依據 POST_CODE_COMPANY 所在縣市套用 107 年度薪資統計，命名為 WORK_AVG_2018
- APPROVE_DATE 在 2021/1/1~2021/12/31，依據 POST_CODE_COMPANY 所在縣市套用 108 年度薪資統計，命名為 WORK_AVG_2019
- APPROVE_DATE 在 2022/1/1~2022/12/31，依據 POST_CODE_COMPANY 所在縣市套用 109 年度薪資統計，命名為 WORK_AVG_2020
- APPROVE_DATE 在 2023/1/1~2023/12/31，依據 POST_CODE_COMPANY 所在縣市套用 110 年度薪資統計，命名為 WORK_AVG_2021

惟因部分外島地區未有相關薪資統計，因此外島部分如金門縣及連江縣採用總計的薪資水準進行取代。與資料集進行 JOIN 後再將 WORK_AVG_XXXX 欄位進行整併，並命名為 WORK_AVG，若 POST_CODE_COMPANY 缺值則同樣給予總計的平均薪資水準。下圖為部分節錄自附錄 5.6 之特徵變數轉換檔的地理區域變數總表：

	A	B	N	O	P	Q
1	ZIP_CODE	ZIP_DESC	WORK_AVG_2018	WORK_AVG_2019	WORK_AVG_2020	WORK_AVG_2021
2	100	臺北市中正區	82.1	84.5	86.3	88.4
3	103	臺北市大同區	82.1	84.5	86.3	88.4
4	104	臺北市中山區	82.1	84.5	86.3	88.4
5	105	臺北市松山區	82.1	84.5	86.3	88.4
6	106	臺北市大安區	82.1	84.5	86.3	88.4
7	108	臺北市萬華區	82.1	84.5	86.3	88.4
8	110	臺北市信義區	82.1	84.5	86.3	88.4
9	111	臺北市士林區	82.1	84.5	86.3	88.4
10	112	臺北市北投區	82.1	84.5	86.3	88.4
11	114	臺北市內湖區	82.1	84.5	86.3	88.4
12	115	臺北市南港區	82.1	84.5	86.3	88.4
13	116	臺北市文山區	82.1	84.5	86.3	88.4
14	200	基隆市仁愛區	60.9	62	62.1	62.9
15	201	基隆市信義區	60.9	62	62.1	62.9
16	202	基隆市中正區	60.9	62	62.1	62.9
17	203	基隆市中山區	60.9	62	62.1	62.9
18	204	基隆市安樂區	60.9	62	62.1	62.9
19	205	基隆市暖暖區	60.9	62	62.1	62.9
20	206	基隆市七堵區	60.9	62	62.1	62.9
21	207	新北市萬里區	61.6	63.3	63.7	65.2
22	208	新北市金山區	61.6	63.3	63.7	65.2
23	220	新北市板橋區	61.6	63.3	63.7	65.2
24	221	新北市汐止區	61.6	63.3	63.7	65.2
25	222	新北市深坑區	61.6	63.3	63.7	65.2
26	223	新北市石碇區	61.6	63.3	63.7	65.2
27	224	新北市瑞芳區	61.6	63.3	63.7	65.2
28	226	新北市平溪區	61.6	63.3	63.7	65.2
29	227	新北市雙溪區	61.6	63.3	63.7	65.2
30	228	新北市貢寮區	61.6	63.3	63.7	65.2

■ 性別倍率

變數名稱為 Gender_Rate，變數來源為資料集所彙整的「性別(gender)」欄位。該變數相關處理是以 3.4 節完成之建模資料集¹⁹裡的三個欄位「平均月薪_全體(MONTH_INCOME_ALL_WA)」、「平均月薪_男(MONTH_INCOME_M_WA)」以及「平均月薪_女(MONTH_INCOME_F_WA)」進行計算即可取得性別倍率變數，計算邏輯請參考以下：

- 若性別(gender)屬於 1，則將平均月薪_男(MONTH_INCOME_M_WA)除以平均月薪_全體(MONTH_INCOME_ALL_WA)作為性別倍率變數值
- 若性別(gender)屬於 2，則將平均月薪_女(MONTH_INCOME_F_WA)除以平均月薪_全體(MONTH_INCOME_ALL_WA)作為性別倍率變數值

■ 無擔保貸款總額度

變數名稱為 VB011_Loan_NonSecur_Limit_NEW_10K_n1，變數來源為 貴行於 2023/9/19 所提供資料表中的「無擔保貸款總額度(VB011_Loan_NonSecur_Limit)」欄位，後續將該欄位重新命名為 VB011_Loan_NonSecur_Limit_NEW，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 VB011_Loan_NonSecur_Limit_NEW 除以 10,000，並且參考同業相關信貸最高可貸額度普遍設定為 500 萬，故設定 500 萬元作為該變數之上限值，即 VB011_Loan_NonSecur_Limit_NEW > 5,000,000，則等於 5,000,000。

■ 近 12 個月平均每發卡銀行永久額度

變數名稱為 VK140_Card_Avg_PerBankLimit_12M_NEW_10K_n1，變數來源為 貴行於 2023/9/19 所提供資料表中的「近 12 個月平均每發卡銀行永久額度(VK140_Card_Avg_PerBankLimit_12M)」欄位，後續將該欄位重新命名為 VK140_Card_Avg_PerBankLimit_12M_NEW，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 VK140_Card_Avg_PerBankLimit_12M_NEW 除以 10,000，並且參考同業相關信用卡給額設定，發卡額度若超過 50 萬元一般屬於頂級卡客，於額度申請時須提供較多元完

¹⁹ 模型 C10~C50 請參考建模資料集-1 名為 ModelTable_Stable_v01

整可認列資產在相當程度以上的證明，故設定 50 萬元作為該變數之上限值，即 $VK140_Card_Avg_PerBankLimit_12M_NEW > 500,000$ ，則等於 500,000。

■ 有擔保貸款高額度註記

變數名稱為 HighLimit_Secur_Mark_C10，變數來源為 貴行於 2023/9/19 所提供資料表中的「擔保貸款總額度(VB002_Loan_Secur_Limit)」欄位，後續將該欄位重新命名為 VB002_Loan_Secur_Limit_NEW，空值則補 0。該變數額外相關處理是檢視並參考變數來源欄位的每個百分位點之數值並反覆測試模型找到最佳之高額度門檻，當變數值大於等於門檻值則賦值為 1，小於門檻值則賦值為 0，而最終 C10 子模型之高額度門檻訂為 3,000,000 元。

■ 近 6 個月高消費註記

變數名稱為 HighSpending_6M_Mark_C10，變數來源為 貴行於 2023/9/19 所提供資料表中的「近 6 個月最高消費金額(VK058_Card_Max_Spending_6M)」欄位，後續將該欄位重新命名為 VK058_Card_Max_Spending_6M_NEW，空值則補 0。該變數額外相關處理是檢視並參考變數來源欄位的每個百分位點之數值並反覆測試模型找到最佳之高消費門檻，當變數值大於等於門檻值則賦值為 1，小於門檻值則賦值為 0，而最終 C10 子模型之高消費門檻訂為 70,000 元。

4.6.5 分量迴歸模型結果

以 C10 子模型投入變數為基礎，進行第 40 百分位數(Q40)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.29412662	-14.85	<0.01
職稱倍率 Jobtitle_Rate	0.12830979	8.21	<0.01
推估總年資註記 SENIORITY_Mark_TOT	0.07215966	4.66	<0.01
工作地年薪平均數 WORK_AVG	0.00336862	7.88	<0.01
性別倍率 Gender_Rate	0.76789666	9.55	<0.01
無擔保貸款總額度	0.00236508	36.05	<0.01

變數名稱	模型係數	T 值	P 值
VB011_Loan_NonSecur_Limit_NEW_10K_n1			
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_10K_n1	0.01348221	9.97	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C10	0.14732292	6.20	<0.01
近 6 個月高消費註記 HighSpending_6M_Mark_C10	0.10690214	4.37	<0.01

以 C10 子模型投入變數為基礎，進行第 35 百分位數(Q35)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.27131074	-13.85	<0.01
職稱倍率 Jobtitle_Rate	0.12367484	10.41	<0.01
推估總年資註記 SENIORITY_Mark_TOT	0.07217284	4.34	<0.01
工作地年薪平均數 WORK_AVG	0.00361193	7.42	<0.01
性別倍率 Gender_Rate	0.70432526	8.09	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00246418	17.86	<0.01
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_10K_n1	0.01280925	8.77	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C10	0.15099995	4.92	<0.01
近 6 個月高消費註記 HighSpending_6M_Mark_C10	0.10069017	4.18	<0.01

以 C10 子模型投入變數為基礎，進行第 30 百分位數(Q30)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.19610566	-13.94	<0.01
職稱倍率 Jobtitle_Rate	0.12239165	8.33	<0.01
推估總年資註記 SENIORITY_Mark_TOT	0.06325369	3.98	<0.01
工作地年薪平均數	0.00324741	6.99	<0.01

變數名稱	模型係數	T 值	P 值
WORK_AVG			
性別倍率 Gender_Rate	0.63280926	7.84	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00240764	15.69	<0.01
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_10K_n1	0.01190098	8.79	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C10	0.13751907	4.90	<0.01
近 6 個月高消費註記 HighSpending_6M_Mark_C10	0.08672949	3.89	<0.01

以 C10 子模型投入變數為基礎，進行第 25 百分位數(Q25)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.15017675	-14.53	<0.01
職稱倍率 Jobtitle_Rate	0.10976260	7.34	<0.01
推估總年資註記 SENIORITY_Mark_TOT	0.05968344	3.59	<0.01
工作地年薪平均數 WORK_AVG	0.00313671	7.50	<0.01
性別倍率 Gender_Rate	0.58104248	7.94	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00237019	24.29	<0.01
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_10K_n1	0.01177758	9.81	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C10	0.14399794	6.30	<0.01
近 6 個月高消費註記 HighSpending_6M_Mark_C10	0.08621120	5.56	<0.01

以 C10 子模型投入變數為基礎，進行第 20 百分位數(Q20)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.11016400	-12.45	<0.01
職稱倍率	0.10464103	9.05	<0.01

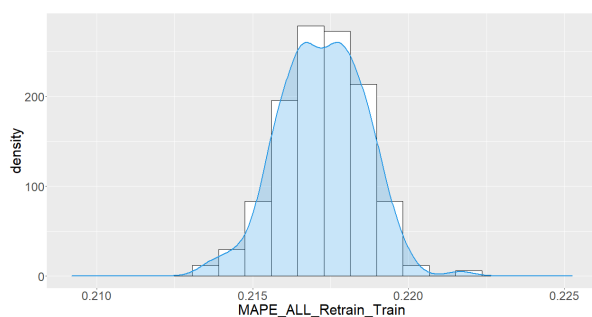
變數名稱	模型係數	T 值	P 值
Jobtitle_Rate			
推估總年資註記 SENIORITY_Mark_TOT	0.03423214	2.16	<0.05
工作地年薪平均數 WORK_AVG	0.00277763	5.99	<0.01
性別倍率 Gender_Rate	0.54315976	6.57	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00246103	19.42	<0.01
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_10K_n1	0.01140911	8.36	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C10	0.13029379	3.62	<0.01
近 6 個月高消費註記 HighSpending_6M_Mark_C10	0.09377342	4.42	<0.01

4.6.6 拔靴法交叉驗證

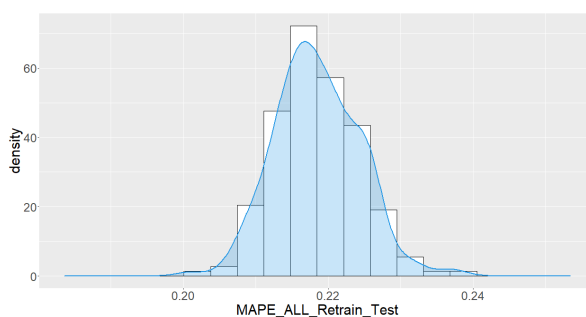
■ 模型效度穩定性

根據交叉驗證重新擬合結果，訓練集與測試集樣本的 MAPE 分布均集中於平均值，且變異係數小於 30%，顯示 C10 子模型的模型效度穩定性佳。

下圖與下表為 200 次 Bootstrapping 交叉驗證的試驗結果，每次試驗將樣本根據核實所得區間進行分層抽樣，並隨機劃分出 80% 為訓練集，20% 為測試集，重新訓練 C10 子模型後所得到樣本整體 MAPE 的分布如下：



80%樣本



20%樣本

樣本範圍	樣本數	MAPE 平均值	MAPE 標準差	MAPE 變異係數
80%樣本	3,205	21.72%	0.13%	0.62%
20%樣本	802	21.85%	0.58%	2.66%

■ 模型係數穩定性

根據交叉驗證重新擬合結果，各模型係數之變異係數均在 30%以下，顯示 C10 子模型的模型係數估計結果具可靠性。

經 200 次 Bootstrapping 交叉驗證，每次試驗抽出 80%的訓練集樣本並重新訓練 C10 子模型後，所得到的模型係數分布如下表：

變數名稱	平均值	標準差	變異係數
職稱倍率 Jobtitle_Rate	0.1588	0.57%	3.62%
推估總年資註記 SENIORITY_Mark_TOT	0.0940	0.68%	7.24%
工作地年薪平均數 WORK_AVG	0.0031	0.02%	4.99%
性別倍率 Gender_Rate	0.7552	3.48%	4.61%
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.0020	0.01%	3.29%
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_10K_n1	0.0157	0.07%	4.34%
有擔保貸款高額度註記 HighLimit_Secur_Mark_C10	0.1539	1.18%	7.64%
近 6 個月高消費註記 HighSpending_6M_Mark_C10	0.1504	0.90%	5.99%

4.7 工程及批發零售業(C20)子模型多變量分析與模型驗證結果

針對子模型 C20(工程及批發零售業)所應用到的資料集說明請參考 4.4 節，而應用到的資料集及結果輸出名稱列示如下：

資料集說明	資料集名稱
訓練集	TrainData_C20_Output.RData
測試集	TestData_C20_Output.RData
建模樣本	ModelData_C20_Output.RData

4.7.1 模型設定

收入距離 = $\alpha + \beta_1$ 職稱倍率 + β_2 推估總年資 + β_3 教育程度年薪中位數 + β_4 性別倍率
+ β_5 工作地年薪平均數 + β_6 無擔保貸款總額度
+ β_7 近 6 個月平均每發卡銀行永久額度 + β_8 有擔保貸款高額度註記
+ β_9 近 6 個月高消費註記 + ε

4.7.2 模型結果

以 3,773 筆訓練集資料配適多元線性迴歸，模型係數四捨五入至小數第 8 位，C20 子模型結果如下：

變數名稱	模型係數	T 值	P 值	VIF 值
截距項	-1.98051158	-20.36	<0.01	-
職稱倍率 Jobtitle_Rate	0.19357928	23.23	<0.01	1.31
推估總年資 SENIORITY_YEAR_TOT	0.00658458	9.94	<0.01	1.32
教育程度年薪中位數 EDU_MID	0.00399162	9.00	<0.01	1.17
性別倍率 Gender_Rate	0.93650892	10.65	<0.01	1.02
工作地年薪平均數 WORK_AVG	0.00336289	7.25	<0.01	1.03

變數名稱	模型係數	T 值	P 值	VIF 值
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00170720	20.82	<0.01	1.19
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01609767	17.28	<0.01	1.56
有擔保貸款高額度註記 HighLimit_Secur_Mark_C20	0.10940416	6.38	<0.01	1.14
近 6 個月高消費註記 HighSpending_6M_Mark_C20	0.12327424	7.50	<0.01	1.22

C20 子模型的 R-squared 值為 0.5676，Adjusted R-squared 值為 0.5666，所有投入變數之模型係數均顯著異於 0。VIF 值均小於 10，變數之間不存在共線性問題。

在業務的應用上，會希望能夠檢視各變數的重要程度，以了解其貢獻性。因此 KPMG 將上述模型進行標準化迴歸，其迴歸係數便能夠揭露其重要程度。結果與變數對應的權重列表如下：

變數名稱	標準化迴歸係數	變數權重
職稱倍率 Jobtitle_Rate	0.2852	21.2%
推估總年資 SENIORITY_YEAR_TOT	0.1224	9.1%
教育程度年薪中位數 EDU_MID	0.1043	7.8%
性別倍率 Gender_Rate	0.1151	8.6%
工作地年薪平均數 WORK_AVG	0.0788	5.9%
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.2436	18.1%
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.2314	17.2%
有擔保貸款高額度註記 HighLimit_Secur_Mark_C20	0.0730	5.4%
近 6 個月高消費註記 HighSpending_6M_Mark_C20	0.0888	6.6%

根據 C20 子模型的模型結果，投入對應的變數即可預測收入距離，後續透過 4.2 節收入距離公式可推導出預測月收入²⁰，結果詳見資料集中的「yhat」欄位。

4.7.3 模型效度

下表為子模型分類 C20 的誤差百分比²¹在各區間的樣本佔比以及 MAPE²²統計結果：

誤差百分比	訓練集	測試集	新增測試集 ²³	全樣本
-100%~-50%	2.65%	3.07%	3.15%	2.73%
-50%~-20%	18.23%	18.75%	13.45%	18.34%
-20%~-10%	12.56%	12.50%	9.28%	12.55%
-10%~0%	12.83%	12.82%	9.36%	12.83%
0%~10%	14.10%	14.72%	10.62%	14.23%
10%~20%	11.87%	11.86%	8.97%	11.87%
20%~40%	16.64%	17.06%	17.39%	16.73%
40%~50%	4.24%	3.28%	7.24%	4.05%
50%~75%	3.79%	2.86%	12.35%	3.60%
75%~100%	1.38%	1.17%	3.70%	1.34%
>100%	1.70%	1.91%	4.48%	1.74%
樣本整體 MAPE	24.77%	24.82%	36.46%	24.78%
樣本數	3,773	944	1,271	4,717

下表則是預估月收入減核實月收入誤差之各區間樣本占比：

誤差範圍	訓練集	測試集	新增測試集	全樣本
<-10 萬	1.46%	1.91%	2.20%	1.55%
-10 萬~-5 萬	2.60%	2.86%	1.89%	2.65%
-5 萬~-3 萬	3.90%	4.56%	3.38%	4.03%
-3 萬~-2 萬	5.54%	4.77%	3.38%	5.38%

²⁰ 依模型估計得到的預測收入距離(\hat{y}_d)會以 $e^{\hat{y}_d + \log(\text{子模型分類下之總體趨勢薪資}[\text{MONTH_INCOME_ALL_WA}])}$ 轉換為預測月收入

²¹ 誤差百分比 = (預測月收入-核實月收入) / 核實月收入

²² 將「yhat」欄位的預測月收入，與「Monthly_Income」欄位的核實月收入套入 MAPE 的計算公式(詳見 2.2.4 節之模型效度指標)，即可計算出預測月收入與核實月收入間的 MAPE

²³ 新增測試集為 2023/12/5 依貴行提出需求新增之測試集，擷取樣本範圍是訓練集以外樣本，並排除(1)非信貸客群 (2)無擔總額度>500 萬之非適用客群 (3)推估年資>40 年之非適用客群 (4)近 6M 每發卡銀行永久額度>50 萬之非適用客群 (5)策略預估大於年收>200 萬之非適用客群 (6)貴行授信政策下不予承做之核實年薪 24 萬以下客群

誤差範圍	訓練集	測試集	新增測試集	全樣本
-2 萬~-1 萬	11.13%	10.91%	8.03%	11.09%
-1 萬~0 萬	21.65%	22.14%	16.37%	21.75%
0 萬~1 萬	34.43%	34.22%	32.02%	34.39%
1 萬~2 萬	12.91%	12.61%	24.70%	12.85%
2 萬~3 萬	2.62%	2.65%	4.17%	2.63%
3 萬~5 萬	2.28%	1.91%	2.52%	2.20%
5 萬~10 萬	1.22%	1.38%	1.34%	1.25%
>10 萬	0.27%	0.11%	2.20%	0.23%
平均絕對誤差(元)	15,501.41	16,104.18	19,533.19	15,622.04

由上表結果顯示，C20 子模型大約有 90%的樣本估計誤差介於 $\pm 50\%$ 之間，而平均絕對誤差²⁴約為 15,501 元。每筆樣本的誤差百分比之詳細結果詳見資料集中的「PerErr」欄位，誤差值詳見「Err」欄位。

■ 各收入區間²⁵

下表是將全樣本分隔成多個預估年收入區間，並計算在各區間樣本下不同誤差範圍下的樣本占比，以及各區間的樣本數、平均絕對誤差(MAE)、平均絕對誤差百分比(MAPE)。後續的應用策略框架報告書將會再說明 C20 子模型與分量迴歸模型的搭配使用。

誤差範圍 (萬元)	預估年收入區間(萬元)												總計
	0-30	30-40	40-50	50-75	75-100	100-150	150-200	200-250	250-300	300-400	400-500	> 500	
<-10	0.0%	0.0%	0.0%	0.4%	0.5%	0.4%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	1.5%
-10~-5	0.0%	0.0%	0.1%	0.7%	0.8%	0.9%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	2.6%
-5~-3	0.0%	0.0%	0.4%	1.6%	1.2%	0.5%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	4.0%
-3~-2	0.0%	0.0%	1.0%	3.0%	1.0%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	5.4%
-2~-1	0.0%	0.2%	3.4%	5.3%	1.6%	0.4%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	11.1%
-1~0	0.0%	1.9%	8.2%	9.5%	1.5%	0.5%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	21.8%
0~1	0.0%	3.3%	15.8%	13.1%	1.3%	0.7%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	34.4%
1~2	0.0%	0.0%	3.1%	7.5%	1.5%	0.5%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	12.8%
2~3	0.0%	0.0%	0.0%	1.2%	0.6%	0.7%	0.1%	0.0%	0.0%	0.1%	0.0%	0.0%	2.6%
3~5	0.0%	0.0%	0.0%	0.1%	1.0%	0.8%	0.2%	0.1%	0.1%	0.1%	0.0%	0.0%	2.2%
5~10	0.0%	0.0%	0.0%	0.0%	0.0%	0.6%	0.3%	0.1%	0.1%	0.1%	0.0%	0.0%	1.3%
>10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%	0.2%

²⁴ 平均絕對誤差(MAE)即為每一筆樣本預測月收入與核實月收入相減後取絕對值再平均

²⁵ 各收入區間是以年收入(ANNINCOME_REVIEW)欄位進行判斷，根據年收入高低劃分成 12 個不同所得級距的區間。

樣本數	0	258	1,512	1,997	511	291	81	29	14	16	5	3	4,717
MAE	-	4,290	7,715	13,038	28,770	41,711	44,287	48,615	63,902	67,438	92,168	203,335	15,622
MAPE	-	13.5%	19.9%	23.6%	33.7%	44.4%	39.1%	49.1%	35.4%	34.5%	39.4%	83.7%	24.8%

4.7.4 模型變數相關說明

■ 職稱倍率

變數名稱為 Jobtitle_Rate，其變數來源由建模資料集與 3.5 節職類別薪資資料處理產出之內外部職稱與職稱倍率彙總表即 TABLE_JOB，利用下述欄位與邏輯進行資料對應與串接而得：

- 職稱資料匹配年月欄位(ADJ_Mapping_YM)，該欄位為透過資料對照年月(Mapping_YM)，將該欄位拆分為資料年度(YEAR)與資料月份(MONTH)並進行調整後合併，其調整邏輯為考量內外部職稱與職稱倍率彙總表的外部資料來源更新時點為每年度的 5 月底至 6 月初，而公布的資料為去年度的 7 月份，舉例如 2022 年 5 月底發布的外部資料為 2021 年 7 月各行業別下各職稱的薪資，因此 2022 年 7 月至 2023 年 6 月的樣本可套用 2021 年 7 月的職稱薪資統計資料，故在建立內外部職稱與職稱倍率彙總表的「職稱資料匹配年月欄位(ADJ_Mapping_YM)」上，當資料對照年月(Mapping_YM)的月份(MONTH)大於等於 7 月時，職稱資料匹配年月(ADJ_Mapping_YM)為資料年度(YEAR)減去 1 後再與字串"07"合併而得，反之當資料對照年月(Mapping_YM)的月份(MONTH)小於 7 月時，職稱資料匹配年月(ADJ_Mapping_YM)為資料年度(YEAR)減去 2 後再與字串"07"合併而得，最後利用上述產出之職稱資料匹配年月(ADJ_Mapping_YM)符合資料年月(DATA_YM)的邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同年度的資料。
- 主計處行業別代碼(CAREER_CODE)，可由 3.2 節之說明以重新歸類後之行業別與邏輯取得，利用該欄位符合行業別代碼大類(CAREER_CODE)邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同行業別代碼的資料。
- 職稱(JOB_CODE)，可由 3.5 節之說明以重新歸類後之職稱別與邏輯取得，利用該欄位符合行內職稱代碼(JOB_TITLE_CODE_BANK)邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同職稱代碼的資料。

針對上述欄位與邏輯利用迴圈方式同時進行判斷，找出並串接每一筆資料集樣本於內外部職稱與職稱倍率彙總表(TABLE_JOB)中相應的職稱倍率_初始(BENCHMARK_RATE)欄位，最後將該欄位命名為 Jobtitle_Rate。

■ 推估總年資

變數名稱為 SENIORITY_Mark_TOT，變數來源為建模資料集所彙整的「年齡(AGE)」與「教育程度(EDUCATION)」欄位，該變數相關處理邏輯參考如下：

- 若教育程度(EDUCATION)屬於 0(研究所以上)者，則將年齡(AGE)減去 24 作為推估總年資(SENIORITY_YEAR_TOT)變數值
- 若教育程度(EDUCATION)屬於 1(大學/大專)者，則將年齡(AGE)減去 22 作為推估總年資(SENIORITY_YEAR_TOT)變數值
- 若教育程度(EDUCATION)屬於 2(高中)者，則將年齡(AGE)減去 18 作為推估總年資(SENIORITY_YEAR_TOT)變數值
- 若教育程度(EDUCATION)屬於 3(國中以下)者，則將年齡(AGE)減去 15 作為推估總年資(SENIORITY_YEAR_TOT)變數值

實際以上述邏輯方式推估，推估總年資會有 1~2 年誤差，舉例如教育程度(EDUCATION)屬於 0(研究所以上)，但年齡(AGE)為 22 者，如遇該情況(於 C20 子模型共 10 筆)則以 0 值做取代。

■ 教育程度年薪中位數

變數名稱為 EDU_MID，變數來源為資料集所彙整的「教育程度(EDUCATION)」欄位。該變數相關處理是以行政院主計總處於 110 年所發布之「工業及服務業受僱員工全年總薪資中位數及分布統計結果」²⁶中的表 1-3「工業及服務業受僱員工全年總薪資中位數」統計之年薪中位數(單位：萬元)作為薪資水準進行套用，資料期間共包含了 105~110 年度，KPMG 也將數據彙整於「特徵變數轉換檔」²⁷的學歷變數總表頁籤中。考量到該份外部資料更新時點為隔年 12 月底，例如：2021 年底發布的外部資料為 2020 年度的薪資統計，因此 2022 年 1 月至 2022 年 12 月的樣本可套用最新發布之 2020 年度的薪資統計，故以下列規則衍生變數：

²⁶ 「工業及服務業受僱員工全年總薪資中位數及分布統計結果」請參考附錄 5.5 之 PDF 檔

²⁷ 「特徵變數轉換檔」的學歷變數總表請參考附錄 5.6 之 EXCEL 檔

- APPROVE_DATE 在 2018/1/1~2018/12/31，依據 EDUCATION 所對應的學歷套用 105 年度薪資統計，命名為 EDU_MID_2016
- APPROVE_DATE 在 2019/1/1~2019/12/31，依據 EDUCATION 所對應的學歷套用 106 年度薪資統計，命名為 EDU_MID_2017
- APPROVE_DATE 在 2020/1/1~2020/12/31，依據 EDUCATION 所對應的學歷套用 107 年度薪資統計，命名為 EDU_MID_2018
- APPROVE_DATE 在 2021/1/1~2021/12/31，依據 EDUCATION 所對應的學歷套用 108 年度薪資統計，命名為 EDU_MID_2019
- APPROVE_DATE 在 2022/1/1~2022/12/31，依據 EDUCATION 所對應的學歷套用 109 年度薪資統計，命名為 EDU_MID_2020
- APPROVE_DATE 在 2023/1/1~2023/12/31，依據 EDUCATION 所對應的學歷套用 110 年度薪資統計，命名為 EDU_MID_2021

與資料集進行 JOIN 後再將 EDU_MID_XXXX 欄位進行整併，並命名為 EDU_MID，若 EDUCATION 缺值則同樣給予總計的中位數薪資水準。下圖為部分節錄自附錄 5.6 之特徵變數轉換檔的學歷變數總表：

	A	B	C	D	E	F	G	H
1	EDU_CODE	EDU_CLASS	EDU_MID_2016	EDU_MID_2017	EDU_MID_2018	EDU_MID_2019	EDU_MID_2020	EDU_MID_2021
2		0 研究所	88.4	90.4	94.1	95.2	96	100.2
3		1 大專	51.1	51.8	53.2	53.6	53.7	54.5
4		2 高中	41.2	42.2	43.6	44.2	44.7	44.9
5		3 國中以下	37.3	38	39.4	40.1	40.5	40.7

■ 性別倍率

變數名稱為 Gender_Rate，變數來源為資料集所彙整的「性別(gender)」欄位。該變數相關處理是以 3.4 節完成之建模資料集²⁸裡的三個欄位「平均月薪_全體(MONTH_INCOME_ALL_WA)」、「平均月薪_男(MONTH_INCOME_M_WA)」以及「平均月薪_女(MONTH_INCOME_F_WA)」進行計算即可取得性別倍率變數，計算邏輯請參考以下：

- 若性別(gender)屬於 1，則將平均月薪_男(MONTH_INCOME_M_WA)除以平均月薪_全體(MONTH_INCOME_ALL_WA)作為性別倍率變數值

²⁸ 模型 C10~C50 請參考建模資料集-1 名為 ModelTable_Stable_v01

- 若性別(gender)屬於 2，則將平均月薪_女(MONTH_INCOME_F_WA)除以平均月薪_全體(MONTH_INCOME_ALL_WA)作為性別倍率變數值

■ 工作地年薪平均數

變數名稱為 WORK_AVG，變數來源為資料集所彙整的「公司郵遞區號(POST_CODE_COMPANY)」欄位。該變數相關處理是以行政院主計總處於 109 發布之「各縣市工業及服務業企業僱用員工薪資統計說明」²⁹中表 1，以及 110 年所發布之「工業及服務業受僱員工全年總薪資中位數及分布統計結果」³⁰中表 2-1「各縣市工業及服務業企業僱用員工全年總薪資」統計之年薪資平均數(單位：萬元)作為薪資水準進行套用，資料期間共包含了 107~110 年度，KPMG 也將其數據彙整於「特徵變數轉換檔」³¹的地理區域變數總表頁籤中。

考量到該份外部資料更新時點為隔年 12 月底，例如：2021 年底發布的外部資料為 2020 年度的薪資統計，因此 2022 年 1 月至 2022 年 12 月的樣本可套用最新發布之 2020 年度的薪資統計，故以下列規則衍生變數：

- APPROVE_DATE 在 2018/01/31~2020/12/31，依據 POST_CODE_COMPANY 所在縣市套用 107 年度薪資統計，命名為 WORK_AVG_2018
- APPROVE_DATE 在 2021/1/1~2021/12/31，依據 POST_CODE_COMPANY 所在縣市套用 108 年度薪資統計，命名為 WORK_AVG_2019
- APPROVE_DATE 在 2022/1/1~2022/12/31，依據 POST_CODE_COMPANY 所在縣市套用 109 年度薪資統計，命名為 WORK_AVG_2020
- APPROVE_DATE 在 2023/1/1~2023/12/31，依據 POST_CODE_COMPANY 所在縣市套用 110 年度薪資統計，命名為 WORK_AVG_2021

惟因部分外島地區未有相關薪資統計，因此外島部分如金門縣及連江縣採用總計的薪資水準進行取代。與資料集進行 JOIN 後再將 WORK_AVG_XXXX 欄位進行整併，並命名為 WORK_AVG，若 POST_CODE_COMPANY 缺值則同樣給予總計的平均薪資水準。下圖為部分節錄自附錄 5.6 之特徵變數轉換檔的地理區域變數總表：

²⁹ 「各縣市工業及服務業企業僱用員工薪資統計說明」請參考附錄 5.4 之 PDF 檔

³⁰ 「工業及服務業受僱員工全年總薪資中位數及分布統計結果」請參考附錄 5.5 之 PDF 檔

³¹ 「特徵變數轉換檔」的地理區域變數總表請參考附錄 5.6 之 EXCEL 檔

	A	B	N	O	P	Q
1	ZIP_CODE	ZIP_DESC	WORK_AVG_2018	WORK_AVG_2019	WORK_AVG_2020	WORK_AVG_2021
2	100	臺北市中正區	82.1	84.5	86.3	88.4
3	103	臺北市大同區	82.1	84.5	86.3	88.4
4	104	臺北市中山區	82.1	84.5	86.3	88.4
5	105	臺北市松山區	82.1	84.5	86.3	88.4
6	106	臺北市大安區	82.1	84.5	86.3	88.4
7	108	臺北市萬華區	82.1	84.5	86.3	88.4
8	110	臺北市信義區	82.1	84.5	86.3	88.4
9	111	臺北市士林區	82.1	84.5	86.3	88.4
10	112	臺北市北投區	82.1	84.5	86.3	88.4
11	114	臺北市內湖區	82.1	84.5	86.3	88.4
12	115	臺北市南港區	82.1	84.5	86.3	88.4
13	116	臺北市文山區	82.1	84.5	86.3	88.4
14	200	基隆市仁愛區	60.9	62	62.1	62.9
15	201	基隆市信義區	60.9	62	62.1	62.9
16	202	基隆市中正區	60.9	62	62.1	62.9
17	203	基隆市中山區	60.9	62	62.1	62.9
18	204	基隆市安樂區	60.9	62	62.1	62.9
19	205	基隆市暖暖區	60.9	62	62.1	62.9
20	206	基隆市七堵區	60.9	62	62.1	62.9
21	207	新北市萬里區	61.6	63.3	63.7	65.2
22	208	新北市金山區	61.6	63.3	63.7	65.2
23	220	新北市板橋區	61.6	63.3	63.7	65.2
24	221	新北市汐止區	61.6	63.3	63.7	65.2
25	222	新北市深坑區	61.6	63.3	63.7	65.2
26	223	新北市石碇區	61.6	63.3	63.7	65.2
27	224	新北市瑞芳區	61.6	63.3	63.7	65.2
28	226	新北市平溪區	61.6	63.3	63.7	65.2
29	227	新北市雙溪區	61.6	63.3	63.7	65.2
30	228	新北市貢寮區	61.6	63.3	63.7	65.2

■ 無擔保貸款總額度

變數名稱為 VB011_Loan_NonSecur_Limit_NEW_10K_n1，變數來源為 貴行於 2023/9/19 所提供資料表中的「無擔保貸款總額度(VB011_Loan_NonSecur_Limit)」欄位，後續將該欄位重新命名為 VB011_Loan_NonSecur_Limit_NEW，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 VB011_Loan_NonSecur_Limit_NEW 除以 10,000，並且參考同業相關信貸最高可貸額度普遍設定為 500 萬，故設定 500 萬元作為該變數之上限值，即 VB011_Loan_NonSecur_Limit_NEW > 5,000,000，則等於 5,000,000。

■ 近 6 個月平均每發卡銀行永久額度

變數名稱為 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1，變數來源為 貴行於 2023/9/19 所提供資料表中的「近 6 個月平均每發卡銀行永久額度(VK138_Card_Avg_PerBankLimit_6M)」欄位，後續將該欄位重新命名為 VK138_Card_Avg_PerBankLimit_6M_NEW，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 VK138_Card_Avg_PerBankLimit_6M_NEW 除以 10,000，並且參考同業相關信用卡給

額設定，發卡額度若超過 50 萬元一般屬於頂級卡客，於額度申請時須提供較多元完整可認列資產在相當程度以上的證明，故設定 50 萬元作為該變數之上限值，即 $VK138_Card_Avg_PerBankLimit_6M_NEW > 500,000$ ，則等於 500,000。

■ 有擔保貸款高額度註記

變數名稱為 HighLimit_Secur_Mark_C20，變數來源為 貴行於 2023/9/19 所提供資料表中的「擔保貸款總額度(VB002_Loan_Secur_Limit)」欄位，後續將該欄位重新命名為 VB002_Loan_Secur_Limit_NEW，空值則補 0。該變數額外相關處理是檢視並參考變數來源欄位的每個百分位點之數值並反覆測試模型找到最佳之高額度門檻，當變數值大於等於門檻值則賦值為 1，小於門檻值則賦值為 0，而最終 C20 子模型之高額度門檻訂為 3,000,000 元。

■ 近 6 個月高消費註記

變數名稱為 HighSpending_6M_Mark_C20，變數來源為 貴行於 2023/9/19 所提供資料表中的「近 6 個月最高消費金額(VK058_Card_Max_Spending_6M)」欄位，後續將該欄位重新命名為 VK058_Card_Max_Spending_6M_NEW，空值則補 0。該變數額外相關處理是檢視並參考變數來源欄位的每個百分位點之數值並反覆測試模型找到最佳之高消費門檻，當變數值大於等於門檻值則賦值為 1，小於門檻值則賦值為 0，而最終 C20 子模型之高消費門檻訂為 90,000 元。

4.7.5 分量迴歸模型結果

以 C20 子模型投入變數為基礎，進行第 40 百分位數(Q40)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-2.10417017	-21.12	<0.01
職稱倍率 Jobtitle_Rate	0.18093293	12.45	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.00556202	7.35	<0.01
教育程度年薪中位數 EDU_MID	0.00399982	7.63	<0.01
性別倍率	0.93876920	10.47	<0.01

變數名稱	模型係數	T 值	P 值
Gender_Rate			
工作地年薪平均數 WORK_AVG	0.00413037	7.93	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00182512	19.66	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01574722	11.97	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C20	0.12957371	7.39	<0.01
近 6 個月高消費註記 HighSpending_6M_Mark_C20	0.11445357	4.41	<0.01

以 C20 子模型投入變數為基礎，進行第 35 百分位數(Q35)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-2.01915123	-20.11	<0.01
職稱倍率 Jobtitle_Rate	0.17391479	13.88	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.00537154	7.36	<0.01
教育程度年薪中位數 EDU_MID	0.00387794	7.18	<0.01
性別倍率 Gender_Rate	0.86718952	9.74	<0.01
工作地年薪平均數 WORK_AVG	0.00380553	7.58	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00191011	39.26	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01445641	11.37	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C20	0.13559477	8.10	<0.01
近 6 個月高消費註記 HighSpending_6M_Mark_C20	0.09104081	3.78	<0.01

以 C20 子模型投入變數為基礎，進行第 30 百分位數(Q30)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-2.01579358	-20.07	<0.01
職稱倍率 Jobtitle_Rate	0.16242927	13.48	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.00525171	7.17	<0.01
教育程度年薪中位數 EDU_MID	0.00398488	7.79	<0.01
性別倍率 Gender_Rate	0.84712978	9.42	<0.01
工作地年薪平均數 WORK_AVG	0.00370073	7.37	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00196027	21.96	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01383129	11.72	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C20	0.15268049	8.34	<0.01
近 6 個月高消費註記 HighSpending_6M_Mark_C20	0.08390354	3.95	<0.01

以 C20 子模型投入變數為基礎，進行第 25 百分位數(Q25)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.96771830	-20.36	<0.01
職稱倍率 Jobtitle_Rate	0.16050357	12.29	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.00503900	7.01	<0.01
教育程度年薪中位數 EDU_MID	0.00384818	6.80	<0.01
性別倍率 Gender_Rate	0.78606031	9.17	<0.01
工作地年薪平均數 WORK_AVG	0.00358014	7.57	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00198596	21.53	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01348049	10.35	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C20	0.14640292	6.49	<0.01

變數名稱	模型係數	T 值	P 值
近 6 個月高消費註記 HighSpending_6M_Mark_C20	0.07708901	3.68	<0.01

以 C20 子模型投入變數為基礎，進行第 20 百分位數(Q20)的分量迴歸配適，模型結果如下：

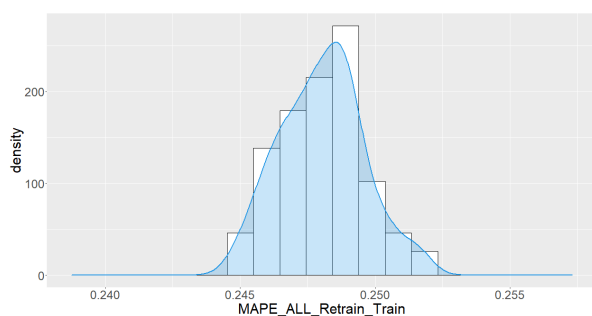
變數名稱	模型係數	T 值	P 值
截距項	-1.95476757	-19.61	<0.01
職稱倍率 Jobtitle_Rate	0.14962186	8.52	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.00442306	5.66	<0.01
教育程度年薪中位數 EDU_MID	0.00370648	6.07	<0.01
性別倍率 Gender_Rate	0.77291373	8.87	<0.01
工作地年薪平均數 WORK_AVG	0.00358718	7.21	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00206356	24.67	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01161244	7.87	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C20	0.13791393	4.93	<0.01
近 6 個月高消費註記 HighSpending_6M_Mark_C20	0.07572567	3.42	<0.01

4.7.6 拔靴法交叉驗證

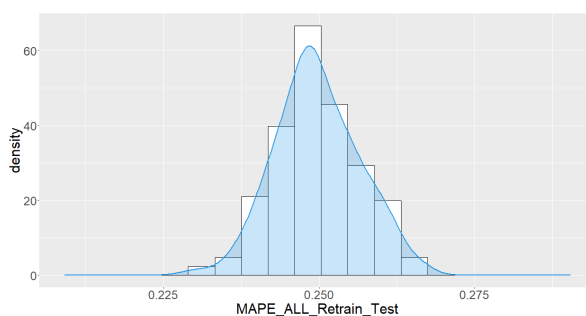
■ 模型效度穩定性

根據交叉驗證重新擬合結果，訓練集與測試集樣本的 MAPE 分布均集中於平均值，且變異係數小於 30%，顯示 C20 子模型的模型效度穩定性佳。

下圖與下表為 200 次 Bootstrapping 交叉驗證的試驗結果，每次試驗將樣本根據核實所得區間進行分層抽樣，並隨機劃分出 80%為訓練集，20%為測試集，重新訓練 C20 子模型後所得到樣本整體 MAPE 的分布如下：



80%樣本



20%樣本

樣本範圍	樣本數	MAPE 平均值	MAPE 標準差	MAPE 變異係數
80%樣本	3,773	24.80%	0.15%	0.62%
20%樣本	944	24.97%	0.68%	2.71%

■ 模型係數穩定性

根據交叉驗證重新擬合結果，各模型係數之變異係數均在 30%以下，顯示 C20 子模型的模型係數估計結果具可靠性。

經 200 次 Bootstrapping 交叉驗證，每次試驗抽出 80%的訓練集樣本並重新訓練 C20 子模型後，所得到的模型係數分布如下表：

變數名稱	平均值	標準差	變異係數
職稱倍率 Jobtitle_Rate	0.1894	0.60%	3.15%
推估總年資 SENIORITY_YEAR_TOT	0.0070	0.03%	4.15%
教育程度年薪中位數 EDU_MID	0.0041	0.02%	5.80%
性別倍率 Gender_Rate	0.9671	3.89%	4.02%
工作地年薪平均數 WORK_AVG	0.0036	0.02%	5.64%
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.0017	0.00%	2.66%
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.0159	0.06%	3.66%
有擔保貸款高額度註記 HighLimit_Secur_Mark_C20	0.0968	0.87%	9.03%
近 6 個月高消費註記 HighSpending_6M_Mark_C20	0.1425	0.92%	6.43%

4.8 技術勞動業(C30)子模型多變量分析與模型驗證結果

針對子模型 C30(技術勞動業)所應用到的資料集說明請參考 4.4 節，而應用到的資料集及結果輸出名稱列示如下：

資料集說明	資料集名稱
訓練集	TrainData_C30_Output.RData
測試集	TestData_C30_Output.RData
建模樣本	ModelData_C30_Output.RData

4.8.1 模型設定

$$\begin{aligned}\text{收入距離} = & \alpha + \beta_1 \text{職稱倍率} + \beta_2 \text{推估總年資} + \beta_3 \text{教育程度年薪中位數} \\ & + \beta_4 \text{工作地年薪平均數} + \beta_5 \text{性別倍率} + \beta_6 \text{上市櫃公司溢酬} \\ & + \beta_7 \text{無擔保貸款總額度} + \beta_8 \text{近 6 個月平均每發卡銀行永久額度} \\ & + \beta_9 \text{有擔保貸款高額度註記} + \beta_{10} \text{近 6 個月高消費註記} + \varepsilon\end{aligned}$$

4.8.2 模型結果

以 11,490 筆訓練集資料配適多元線性迴歸，模型係數四捨五入至小數第 8 位，C30 子模型結果如下：

變數名稱	模型係數	T 值	P 值	VIF 值
截距項	-1.74074476	-55.16	<0.01	-
職稱倍率 Jobtitle_Rate	0.14915013	38.91	<0.01	1.40
推估總年資 SENIORITY_YEAR_TOT	0.00799479	20.07	<0.01	1.26
教育程度年薪中位數 EDU_MID	0.00450699	22.89	<0.01	1.38
工作地年薪平均數 WORK_AVG	0.00353685	13.85	<0.01	1.10
性別倍率 Gender_Rate	0.54080319	22.99	<0.01	1.03
上市櫃公司溢酬 Salary_Company_Prem02	0.15996926	30.34	<0.01	1.13

變數名稱	模型係數	T 值	P 值	VIF 值
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00127535	36.31	<0.01	1.24
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01669597	34.99	<0.01	1.72
有擔保貸款高額度註記 HighLimit_Secur_Mark_C30	0.09302816	11.35	<0.01	1.16
近 6 個月高消費註記 HighSpending_6M_Mark_C30	0.13729937	14.09	<0.01	1.27

C30 子模型的 R-squared 值為 0.6656，Adjusted R-squared 值為 0.6653，所有投入變數之模型係數均顯著異於 0。VIF 值均小於 10，變數之間不存在共線性問題。

在業務的應用上，會希望能夠檢視各變數的重要程度，以了解其貢獻性。因此 KPMG 將上述模型進行標準化迴歸，其迴歸係數便能夠揭露其重要程度。結果與變數對應的權重列表如下：

變數名稱	標準化迴歸係數	變數權重
職稱倍率 Jobtitle_Rate	0.2486	16.5%
推估總年資 SENIORITY_YEAR_TOT	0.1218	8.1%
教育程度年薪中位數 EDU_MID	0.1451	9.6%
工作地年薪平均數 WORK_AVG	0.0786	5.2%
性別倍率 Gender_Rate	0.1258	8.3%
上市櫃公司溢酬 Salary_Company_Prem02	0.1739	11.5%
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.2179	14.4%
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.2474	16.4%
有擔保貸款高額度註記 HighLimit_Secur_Mark_C30	0.0659	4.4%

變數名稱	標準化迴歸係數	變數權重
近 6 個月高消費註記 HighSpending_6M_Mark_C30	0.0857	5.7%

根據 C30 子模型的模型結果，投入對應的變數即可預測收入距離，後續透過 4.2 節收入距離公式可推導出預測月收入³²，結果詳見資料集中的「yhat」欄位。

4.8.3 模型效度

下表為子模型分類 C30 的誤差百分比³³在各區間的樣本佔比以及 MAPE³⁴統計結果：

誤差百分比	訓練集	測試集	新增測試集 ³⁵	全樣本
-100%~-50%	2.52%	2.44%	2.79%	2.51%
-50%~-20%	18.84%	20.08%	18.08%	19.09%
-20%~-10%	12.05%	12.39%	11.36%	12.11%
-10%~0%	13.60%	13.23%	12.05%	13.53%
0%~10%	13.58%	13.99%	12.69%	13.66%
10%~20%	11.34%	10.76%	9.68%	11.22%
20%~40%	15.86%	15.07%	13.86%	15.70%
40%~50%	4.53%	4.56%	4.77%	4.54%
50%~75%	5.16%	5.05%	9.03%	5.14%
75%~100%	1.49%	1.36%	3.67%	1.46%
>100%	1.03%	1.08%	2.01%	1.04%
樣本整體 MAPE	24.64%	24.41%	29.39%	24.59%
樣本數	11,490	2,873	3,080	14,363

下表則是預估月收入減核實月收入誤差之各區間樣本占比：

³² 依模型估計得到的預測收入距離(\hat{y}_a)會以 $e^{\hat{y}_a + \log(\text{子模型分類下之總體趨勢薪資}[\text{MONTH_INCOME_ALL_WA}])}$ 轉換為預測月收入

³³ 誤差百分比 = (預測月收入-核實月收入) / 核實月收入

³⁴ 將「yhat」欄位的預測月收入，與「Monthly_Income」欄位的核實月收入套入 MAPE 的計算公式(詳見 2.2.4 節之模型效度指標)，即可計算出預測月收入與核實月收入間的 MAPE

³⁵ 新增測試集為 2023/12/5 依貴行提出需求新增之測試集，擷取樣本範圍是訓練集以外樣本，並排除(1)非信貸客群 (2)無擔總額度>500 萬之非適用客群 (3)近 6M 每發卡銀行永久額度>50 萬之非適用客群 (4)策略預估大於年收>200 萬之非適用客群 (5)貴行授信政策下不予承做之核實年薪 24 萬以下客群

誤差範圍	訓練集	測試集	新增測試集	全樣本
<-10 萬	1.98%	2.05%	2.18%	1.99%
-10 萬~-5 萬	4.26%	3.90%	3.21%	4.19%
-5 萬~-3 萬	4.79%	5.57%	5.03%	4.94%
-3 萬~-2 萬	6.00%	5.74%	5.32%	5.95%
-2 萬~-1 萬	10.26%	11.21%	10.32%	10.45%
-1 萬~0 萬	19.73%	19.67%	18.21%	19.72%
0 萬~1 萬	27.38%	26.56%	25.39%	27.22%
1 萬~2 萬	15.37%	15.38%	20.97%	15.37%
2 萬~3 萬	5.08%	4.49%	5.94%	4.96%
3 萬~5 萬	2.80%	3.03%	2.76%	2.85%
5 萬~10 萬	1.59%	1.50%	0.62%	1.57%
>10 萬	0.76%	0.90%	0.03%	0.79%
平均絕對誤差(元)	19,907.59	20,163.07	20,114.61	19,958.69

由上表結果顯示，C30 子模型大約有 91% 的樣本估計誤差介於 $\pm 50\%$ 之間，而平均絕對誤差³⁶約為 19,908 元。每筆樣本的誤差百分比之詳細結果詳見資料集中的「PerErr」欄位，誤差值詳見「Err」欄位。

■ 各收入區間³⁷

下表是將全樣本分隔成多個預估年收入區間，並計算在各區間樣本下不同誤差範圍下的樣本占比，以及各區間的樣本數、平均絕對誤差(MAE)、平均絕對誤差百分比(MAPE)。後續的應用策略框架報告書將會再說明 C10 子模型與分量迴歸模型的搭配使用。

誤差範圍 (萬元)	預估年收入區間(萬元)												總計
	0-30	30-40	40-50	50-75	75-100	100-150	150-200	200-250	250-300	300-400	400-500	> 500	
<-10	0.0%	0.0%	0.0%	0.1%	0.3%	0.7%	0.4%	0.3%	0.1%	0.0%	0.0%	0.0%	2.0%
-10~-5	0.0%	0.0%	0.0%	0.6%	1.1%	1.5%	0.5%	0.2%	0.2%	0.1%	0.0%	0.0%	4.2%
-5~-3	0.0%	0.0%	0.2%	1.8%	1.4%	1.0%	0.4%	0.1%	0.0%	0.0%	0.0%	0.0%	4.9%
-3~-2	0.0%	0.0%	0.4%	2.9%	1.5%	0.7%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	5.9%
-2~-1	0.0%	0.1%	1.3%	5.4%	2.0%	1.2%	0.3%	0.1%	0.1%	0.0%	0.0%	0.0%	10.5%
-1~0	0.0%	0.2%	3.8%	10.9%	2.9%	1.4%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	19.7%

³⁶ 平均絕對誤差(MAE)即為每一筆樣本預測月收入與核實月收入相減後取絕對值再平均

³⁷ 各收入區間是以年收入(ANNINCOME_REVIEW)欄位進行判斷，根據年收入高低劃分成 12 個不同所得級距的區間。

0~1	0.0%	0.5%	6.2%	15.1%	3.4%	1.7%	0.3%	0.1%	0.0%	0.1%	0.0%	0.0%	27.2%
1~2	0.0%	0.0%	1.5%	8.8%	3.0%	1.5%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	15.4%
2~3	0.0%	0.0%	0.0%	1.5%	1.8%	1.2%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	5.0%
3~5	0.0%	0.0%	0.0%	0.1%	0.5%	1.2%	0.5%	0.3%	0.1%	0.1%	0.0%	0.0%	2.8%
5~10	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%	0.4%	0.3%	0.2%	0.2%	0.0%	0.0%	1.6%
>10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.2%	0.2%	0.3%	0.8%
樣本數	0	117	1,931	6,782	2,576	1,760	568	270	139	125	47	48	14,363
MAE	-	5,710	7,637	11,839	21,290	33,346	48,356	64,933	66,239	68,230	101,835	206,415	19,959
MAPE	-	16.7%	19.7%	22.4%	26.1%	28.7%	32.3%	35.6%	37.6%	34.4%	49.8%	74.1%	24.6%

4.8.4 模型變數相關說明

■ 職稱倍率

變數名稱為 Jobtitle_Rate，其變數來源由建模資料集與 3.5 節職類別薪資資料處理產出之內外部職稱與職稱倍率彙總表即 TABLE_JOB，利用下述欄位與邏輯進行資料對應與串接而得：

- 職稱資料匹配年月欄位(ADJ_Mapping_YM)，該欄位為透過資料對照年月(Mapping_YM)，將該欄位拆分為資料年度(YEAR)與資料月份(MONTH)並進行調整後合併，其調整邏輯為考量內外部職稱與職稱倍率彙總表的外部資料來源更新時點為每年度的 5 月底至 6 月初，而公布的資料為去年度的 7 月份，舉例如 2022 年 5 月底發布的外部資料為 2021 年 7 月各行業別下各職稱的薪資，因此 2022 年 7 月至 2023 年 6 月的樣本可套用 2021 年 7 月的職稱薪資統計資料，故在建立內外部職稱與職稱倍率彙總表的「職稱資料匹配年月欄位(ADJ_Mapping_YM)」上，當資料對照年月(Mapping_YM)的月份(MONTH)大於等於 7 月時，職稱資料匹配年月(ADJ_Mapping_YM)為資料年度(YEAR)減去 1 後再與字串"07"合併而得，反之當資料對照年月(Mapping_YM)的月份(MONTH)小於 7 月時，職稱資料匹配年月(ADJ_Mapping_YM)為資料年度(YEAR)減去 2 後再與字串"07"合併而得，最後利用上述產出之職稱資料匹配年月(ADJ_Mapping_YM)符合資料年月(DATA_YM)的邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同年度的資料。
- 主計處行業別代碼(CAREER_CODE)，可由 3.2 節之說明以重新歸類後之行業別與邏輯取得，利用該欄位符合行業別代碼大類(CAREER_CODE)邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同行業別代碼的資料。

- 職稱(JOB_CODE)·可由 3.5 節之說明以重新歸類後之職稱別與邏輯取得·利用該欄位符合行內職稱代碼(JOB_TITLE_CODE_BANK) 邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同職稱代碼的資料。

針對上述欄位與邏輯利用迴圈方式同時進行判斷·找出並串接每一筆資料集樣本於內外部職稱與職稱倍率彙總表(TABLE_JOB)中相應的職稱倍率_初始(BENCHMARK_RATE)欄位·最後將該欄位命名為 Jobtitle_Rate。

■ 推估總年資

變數名稱為 SENIORITY_Mark_TOT·變數來源為建模資料集所彙整的「年齡(AGE)」與「教育程度(EDUCATION)」欄位·該變數相關處理邏輯參考如下：

- 若教育程度(EDUCATION)屬於 0(研究所以上)者·則將年齡(AGE)減去 24 作為推估總年資(SENIORITY_YEAR_TOT)變數值
- 若教育程度(EDUCATION)屬於 1(大學/大專)者·則將年齡(AGE)減去 22 作為推估總年資(SENIORITY_YEAR_TOT)變數值
- 若教育程度(EDUCATION)屬於 2(高中)者·則將年齡(AGE)減去 18 作為推估總年資(SENIORITY_YEAR_TOT)變數值
- 若教育程度(EDUCATION)屬於 3(國中以下)者·則將年齡(AGE)減去 15 作為推估總年資(SENIORITY_YEAR_TOT)變數值

實際以上述邏輯方式推估·推估總年資會有 1~2 年誤差·舉例如教育程度(EDUCATION)屬於 0(研究所以上)·但年齡(AGE)為 22 者·如遇該情況(於 C30 子模型共 11 筆)則以 0 值做取代。

■ 教育程度年薪中位數

變數名稱為 EDU_MID·變數來源為資料集所彙整的「教育程度(EDUCATION)」欄位。該變數相關處理是以行政院主計總處於 110 年所發布之「工業及服務業受僱員工全年總薪資中位數及分布統計結果」³⁸中的表 1-3「工業及服務業受僱員工全年總薪資中位數」統計之年薪中位數(單位：萬元)作為薪資水準進行套用·資料期間共包含了 105~110 年度·KPMG 也將數據彙整於「特徵變數轉換檔」³⁹的學歷變數總表頁籤中。

³⁸ 「工業及服務業受僱員工全年總薪資中位數及分布統計結果」請參考附錄 5.5 之 PDF 檔

³⁹ 「特徵變數轉換檔」的學歷變數總表請參考附錄 5.6 之 EXCEL 檔

考量到該份外部資料更新時點為隔年 12 月底，例如：2021 年底發布的外部資料為 2020 年度的薪資統計，因此 2022 年 1 月至 2022 年 12 月的樣本可套用最新發布之 2020 年度的薪資統計，故以下列規則衍生變數：

- APPROVE_DATE 在 2018/1/1~2018/12/31，依據 EDUCATION 所對應的學歷套用 105 年度薪資統計，命名為 EDU_MID_2016
- APPROVE_DATE 在 2019/1/1~2019/12/31，依據 EDUCATION 所對應的學歷套用 106 年度薪資統計，命名為 EDU_MID_2017
- APPROVE_DATE 在 2020/1/1~2020/12/31，依據 EDUCATION 所對應的學歷套用 107 年度薪資統計，命名為 EDU_MID_2018
- APPROVE_DATE 在 2021/1/1~2021/12/31，依據 EDUCATION 所對應的學歷套用 108 年度薪資統計，命名為 EDU_MID_2019
- APPROVE_DATE 在 2022/1/1~2022/12/31，依據 EDUCATION 所對應的學歷套用 109 年度薪資統計，命名為 EDU_MID_2020
- APPROVE_DATE 在 2023/1/1~2023/12/31，依據 EDUCATION 所對應的學歷套用 110 年度薪資統計，命名為 EDU_MID_2021

與資料集進行 JOIN 後再將 EDU_MID_XXXX 欄位進行整併，並命名為 EDU_MID，若 EDUCATION 缺值則同樣給予總計的中位數薪資水準。下圖為部分節錄自附錄 5.6 之特徵變數轉換檔的學歷變數總表：

	A	B	C	D	E	F	G	H
1	EDU_CODE	EDU_CLASS	EDU_MID_2016	EDU_MID_2017	EDU_MID_2018	EDU_MID_2019	EDU_MID_2020	EDU_MID_2021
2		0 研究所	88.4	90.4	94.1	95.2	96	100.2
3		1 大專	51.1	51.8	53.2	53.6	53.7	54.5
4		2 高中	41.2	42.2	43.6	44.2	44.7	44.9
5		3 國中以下	37.3	38	39.4	40.1	40.5	40.7

■ 工作地年薪平均數

變數名稱為 WORK_AVG，變數來源為資料集所彙整的「公司郵遞區號 (POST_CODE_COMPANY)」欄位。該變數相關處理是以行政院主計總處於 109 發布之「各縣市工業及服務業企業僱用員工薪資統計說明」⁴⁰中表 1，以及 110 年所發布之「工業及服務業受僱員工全年總薪資中位數及分布統計結果」⁴¹中表 2-1「各縣市工業

⁴⁰ 「各縣市工業及服務業企業僱用員工薪資統計說明」請參考附錄 5.4 之 PDF 檔

⁴¹ 「工業及服務業受僱員工全年總薪資中位數及分布統計結果」請參考附錄 5.5 之 PDF 檔

及服務業企業僱用員工全年總薪資」統計之年薪資平均數(單位：萬元)作為薪資水準進行套用，資料期間共包含了 107~110 年度，KPMG 也將其數據彙整於「特徵變數轉換檔」⁴²的地理區域變數總表頁籤中。

考量到該份外部資料更新時點為隔年 12 月底，例如：2021 年底發布的外部資料為 2020 年度的薪資統計，因此 2022 年 1 月至 2022 年 12 月的樣本可套用最新發布之 2020 年度的薪資統計，故以下列規則衍生變數：

- APPROVE_DATE 在 2018/01/31~2020/12/31，依據 POST_CODE_COMPANY 所在縣市套用 107 年度薪資統計，命名為 WORK_AVG_2018
- APPROVE_DATE 在 2021/1/1~2021/12/31，依據 POST_CODE_COMPANY 所在縣市套用 108 年度薪資統計，命名為 WORK_AVG_2019
- APPROVE_DATE 在 2022/1/1~2022/12/31，依據 POST_CODE_COMPANY 所在縣市套用 109 年度薪資統計，命名為 WORK_AVG_2020
- APPROVE_DATE 在 2023/1/1~2023/12/31，依據 POST_CODE_COMPANY 所在縣市套用 110 年度薪資統計，命名為 WORK_AVG_2021

惟因部分外島地區未有相關薪資統計，因此外島部分如金門縣及連江縣採用總計的薪資水準進行取代。與資料集進行 JOIN 後再將 WORK_AVG_XXXX 欄位進行整併，並命名為 WORK_AVG，若 POST_CODE_COMPANY 缺值則同樣給予總計的平均薪資水準。下圖為部分節錄自附錄 5.6 之特徵變數轉換檔的地理區域變數總表：

⁴² 「特徵變數轉換檔」的地理區域變數總表請參考附錄 5.6 之 EXCEL 檔

	A	B	N	O	P	Q
1	ZIP_CODE	ZIP_DESC	WORK_AVG_2018	WORK_AVG_2019	WORK_AVG_2020	WORK_AVG_2021
2	100	臺北市中正區	82.1	84.5	86.3	88.4
3	103	臺北市大同區	82.1	84.5	86.3	88.4
4	104	臺北市中山區	82.1	84.5	86.3	88.4
5	105	臺北市松山區	82.1	84.5	86.3	88.4
6	106	臺北市大安區	82.1	84.5	86.3	88.4
7	108	臺北市萬華區	82.1	84.5	86.3	88.4
8	110	臺北市信義區	82.1	84.5	86.3	88.4
9	111	臺北市士林區	82.1	84.5	86.3	88.4
10	112	臺北市北投區	82.1	84.5	86.3	88.4
11	114	臺北市內湖區	82.1	84.5	86.3	88.4
12	115	臺北市南港區	82.1	84.5	86.3	88.4
13	116	臺北市文山區	82.1	84.5	86.3	88.4
14	200	基隆市仁愛區	60.9	62	62.1	62.9
15	201	基隆市信義區	60.9	62	62.1	62.9
16	202	基隆市中正區	60.9	62	62.1	62.9
17	203	基隆市中山區	60.9	62	62.1	62.9
18	204	基隆市安樂區	60.9	62	62.1	62.9
19	205	基隆市暖暖區	60.9	62	62.1	62.9
20	206	基隆市七堵區	60.9	62	62.1	62.9
21	207	新北市萬里區	61.6	63.3	63.7	65.2
22	208	新北市金山區	61.6	63.3	63.7	65.2
23	220	新北市板橋區	61.6	63.3	63.7	65.2
24	221	新北市汐止區	61.6	63.3	63.7	65.2
25	222	新北市深坑區	61.6	63.3	63.7	65.2
26	223	新北市石碇區	61.6	63.3	63.7	65.2
27	224	新北市瑞芳區	61.6	63.3	63.7	65.2
28	226	新北市平溪區	61.6	63.3	63.7	65.2
29	227	新北市雙溪區	61.6	63.3	63.7	65.2
30	228	新北市貢寮區	61.6	63.3	63.7	65.2

■ 性別倍率

變數名稱為 **Gender_Rate**，變數來源為資料集所彙整的「性別(gender)」欄位。該變數相關處理是以 3.4 節完成之建模資料集⁴³裡的三個欄位「平均月薪_全體(MONTH_INCOME_ALL_WA)」、「平均月薪_男(MONTH_INCOME_M_WA)」以及「平均月薪_女(MONTH_INCOME_F_WA)」進行計算即可取得性別倍率變數，計算邏輯請參考以下：

- 若性別(gender)屬於 1，則將平均月薪_男(MONTH_INCOME_M_WA)除以平均月薪_全體(MONTH_INCOME_ALL_WA)作為性別倍率變數值
- 若性別(gender)屬於 2，則將平均月薪_女(MONTH_INCOME_F_WA)除以平均月薪_全體(MONTH_INCOME_ALL_WA)作為性別倍率變數值

⁴³ 模型 C10~C50 請參考建模資料集-1 名為 ModelTable_Stable_v01

■ 上市櫃公司溢酬

變數名稱為 `Salary_Company_Prem02`，變數來源為資料集所彙整的「公司名稱(COMPANY)」、「主計處薪資水準(MONTH_INCOME_CAREERCODE)」欄位。

首先，須將資料處理篇所整理的「上市櫃公司名單列表」與資料集進行 JOIN，若資料集公司名稱(COMPANY)欄位中含有「上市櫃公司名單列表」所列公司名稱關鍵字串者，則賦予該樣本對應的公司代碼(Symbol)，惟派遣、工讀生等職位之敘薪與其他正職員工相異甚大，其薪資水準不宜視同上市櫃公司之正職員工，故當 `OC_CODE_REVIEW` 為 28, 29, 30, 31, 32, 33, 34, 35, 37, 55, 63, 67, 68 時，公司代碼(Symbol)欄位為空。

其次，再依據公司代碼(Symbol)以及資料對照年月(Mapping_YM)，與「上市櫃員工薪資對照表」進行匹配，即當公司代碼(Symbol)相同且資料對照年月(Mapping_YM)介於套用起點與套用終點時，給予「非擔任主管職務之全時員工資訊_員工薪資_平均數_仟元每人(Salary_Company)」欄位的值作為其所屬公司的薪資水準，並將其乘以 1,000 再除以 12 轉換為平均月收入(單位：元)。

最後，設定一個新的 `Salary_Company_Prem` 欄位，當該客戶為上市櫃公司正職員工，即 `Salary_Company` 非空值時，則賦值為 `Salary_Company` 當作該公司薪資水準；反之，當該客戶非屬上市櫃公司正職員工時，則賦值為該行業的主計處薪資水準(MONTH_INCOME_CAREERCODE)作為其所屬產業的薪資水準。最終再將 `Salary_Company_Prem` 除以主計處薪資水準(MONTH_INCOME_CAREERCODE)，即可得到上市櫃公司溢酬。

■ 無擔保貸款總額度

變數名稱為 `VB011_Loan_NonSecur_Limit_NEW_10K_n1`，變數來源為貴行於 2023/9/19 所提供資料表中的「無擔保貸款總額度(VB011_Loan_NonSecur_Limit)」欄位，後續將該欄位重新命名為 `VB011_Loan_NonSecur_Limit_NEW`，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 `VB011_Loan_NonSecur_Limit_NEW` 除以 10,000，並且參考同業相關信貸最高可貸額度普遍設定為 500 萬，故設定 500 萬元作為該變數之上限值，即 `VB011_Loan_NonSecur_Limit_NEW > 5,000,000`，則等於 5,000,000。

■ 近 6 個月平均每發卡銀行永久額度

變數名稱為 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1，變數來源為 貴行於 2023/9/19 所提供資料表中的「近 6 個月平均每發卡銀行永久額度 (VK138_Card_Avg_PerBankLimit_6M)」欄位，後續將該欄位重新命名為 VK138_Card_Avg_PerBankLimit_6M_NEW，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 VK138_Card_Avg_PerBankLimit_6M_NEW 除以 10,000，並且參考同業相關信用卡給額設定，發卡額度若超過 50 萬元一般屬於頂級卡客，於額度申請時須提供較多元完整可認列資產在相當程度以上的證明，故設定 50 萬元作為該變數之上限值，即 $VK138_Card_Avg_PerBankLimit_6M_NEW > 500,000$ ，則等於 500,000。

■ 有擔保貸款高額度註記

變數名稱為 HighLimit_Secur_Mark_C30，變數來源為 貴行於 2023/9/19 所提供資料表中的「擔保貸款總額度(VB002_Loan_Secur_Limit)」欄位，後續將該欄位重新命名為 VB002_Loan_Secur_Limit_NEW，空值則補 0。該變數額外相關處理是檢視並參考變數來源欄位的每個百分位點之數值並反覆測試模型找到最佳之高額度門檻，當變數值大於等於門檻值則賦值為 1，小於門檻值則賦值為 0，而最終 C30 子模型之高額度門檻訂為 3,000,000 元。

■ 近 6 個月高消費註記

變數名稱為 HighSpending_6M_Mark_C30，變數來源為 貴行於 2023/9/19 所提供資料表中的「近 6 個月最高消費金額(VK058_Card_Max_Spending_6M)」欄位，後續將該欄位重新命名為 VK058_Card_Max_Spending_6M_NEW，空值則補 0。該變數額外相關處理是檢視並參考變數來源欄位的每個百分位點之數值並反覆測試模型找到最佳之高消費門檻，當變數值大於等於門檻值則賦值為 1，小於門檻值則賦值為 0，而最終 C30 子模型之高消費門檻訂為 100,000 元。

4.8.5 分量迴歸模型結果

以 C30 子模型投入變數為基礎，進行第 40 百分位數(Q40)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.81183735	-53.96	<0.01
職稱倍率 Jobtitle_Rate	0.14569594	29.36	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.00741299	16.58	<0.01
教育程度年薪中位數 EDU_MID	0.00417863	19.36	<0.01
工作地年薪平均數 WORK_AVG	0.00322771	11.39	<0.01
性別倍率 Gender_Rate	0.57249439	23.02	<0.01
上市櫃公司溢酬 Salary_Company_Prem02	0.16093024	29.22	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00143270	35.81	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01575790	26.74	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C30	0.09575212	9.67	<0.01
近 6 個月高消費註記 HighSpending_6M_Mark_C30	0.11772282	9.30	<0.01

以 C30 子模型投入變數為基礎，進行第 35 百分位數(Q35)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.82316164	-52.89	<0.01
職稱倍率 Jobtitle_Rate	0.14438621	26.60	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.00701119	14.95	<0.01
教育程度年薪中位數 EDU_MID	0.00433321	19.76	<0.01
工作地年薪平均數 WORK_AVG	0.00305413	10.67	<0.01
性別倍率 Gender_Rate	0.56264439	22.11	<0.01
上市櫃公司溢酬 Salary_Company_Prem02	0.15972970	23.26	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00147245	32.09	<0.01
近 6 個月平均每發卡銀行永久額度	0.01541852	25.12	<0.01

變數名稱	模型係數	T 值	P 值
VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1			
有擔保貸款高額度註記 HighLimit_Secur_Mark_C30	0.08979864	8.62	<0.01
近 6 個月高消費註記 HighSpending_6M_Mark_C30	0.11052354	8.03	<0.01

以 C30 子模型投入變數為基礎，進行第 30 百分位數(Q30)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.83151108	-52.04	<0.01
職稱倍率 Jobtitle_Rate	0.14191128	23.16	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.00633016	13.22	<0.01
教育程度年薪中位數 EDU_MID	0.00425245	16.88	<0.01
工作地年薪平均數 WORK_AVG	0.00315584	10.67	<0.01
性別倍率 Gender_Rate	0.55357824	21.70	<0.01
上市櫃公司溢酬 Salary_Company_Prem02	0.15437497	22.00	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00145051	27.71	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01555446	26.50	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C30	0.08295687	7.82	<0.01
近 6 個月高消費註記 HighSpending_6M_Mark_C30	0.10135085	7.26	<0.01

以 C30 子模型投入變數為基礎，進行第 25 百分位數(Q25)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.83348431	-49.43	<0.01
職稱倍率 Jobtitle_Rate	0.13547320	22.37	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.00599597	12.53	<0.01

變數名稱	模型係數	T 值	P 值
教育程度年薪中位數 EDU_MID	0.00413710	16.31	<0.01
工作地年薪平均數 WORK_AVG	0.00292489	8.90	<0.01
性別倍率 Gender_Rate	0.54448875	20.49	<0.01
上市櫃公司溢酬 Salary_Company_Prem02	0.15585862	22.33	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00144062	28.68	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01575055	23.69	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C30	0.08829569	8.72	<0.01
近 6 個月高消費註記 HighSpending_6M_Mark_C30	0.08961011	6.66	<0.01

以 C30 子模型投入變數為基礎，進行第 20 百分位數(Q20)的分量迴歸配適，模型結果如下：

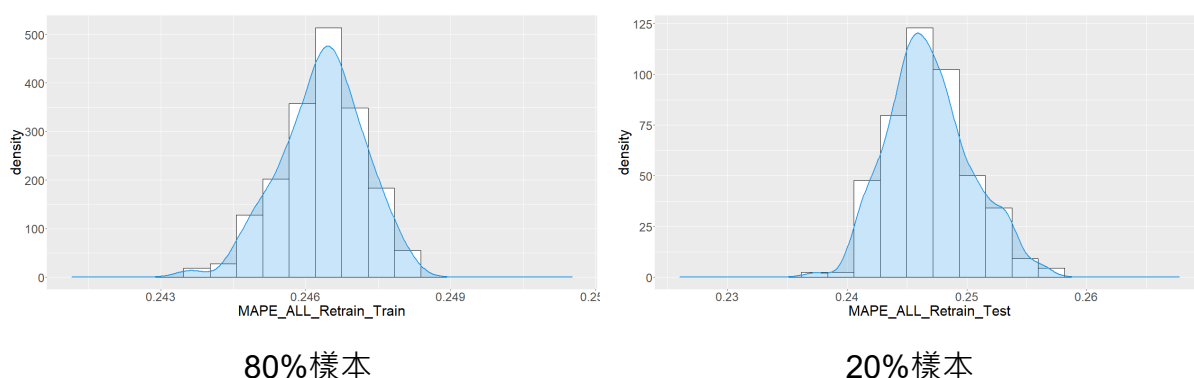
變數名稱	模型係數	T 值	P 值
截距項	-1.84515319	-54.98	<0.01
職稱倍率 Jobtitle_Rate	0.12934530	22.38	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.00583085	12.35	<0.01
教育程度年薪中位數 EDU_MID	0.00420980	17.13	<0.01
工作地年薪平均數 WORK_AVG	0.00287368	9.71	<0.01
性別倍率 Gender_Rate	0.53228632	22.10	<0.01
上市櫃公司溢酬 Salary_Company_Prem02	0.15163714	22.25	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00144401	27.50	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01549795	24.99	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C30	0.08673541	9.00	<0.01
近 6 個月高消費註記 HighSpending_6M_Mark_C30	0.08224388	5.90	<0.01

4.8.6 拔靴法交叉驗證

■ 模型效度穩定性

根據交叉驗證重新擬合結果，訓練集與測試集樣本的 **MAPE** 分布均集中於平均值，且變異係數小於 30%，顯示 C30 子模型的模型效度穩定性佳。

下圖與下表為 200 次 Bootstrapping 交叉驗證的試驗結果，每次試驗將樣本根據核實所得區間進行分層抽樣，並隨機劃分出 80% 為訓練集，20% 為測試集，重新訓練 C30 子模型後所得到樣本整體 **MAPE** 的分布如下：



樣本範圍	樣本數	MAPE 平均值	MAPE 標準差	MAPE 變異係數
80%樣本	11,490	24.63%	0.09%	0.35%
20%樣本	2,873	24.69%	0.35%	1.41%

■ 模型係數穩定性

根據交叉驗證重新擬合結果，各模型係數之變異係數均在 30% 以下，顯示 C30 子模型的模型係數估計結果具可靠性。

經 200 次 Bootstrapping 交叉驗證，每次試驗抽出 80% 的訓練集樣本並重新訓練 C30 子模型後，所得到的模型係數分布如下表：

變數名稱	平均值	標準差	變異係數
職稱倍率 Jobtitle_Rate	0.1511	0.24%	1.60%
推估總年資 SENIORITY_YEAR_TOT	0.0080	0.02%	2.31%
教育程度年薪中位數 EDU_MID	0.0046	0.01%	2.10%

變數名稱	平均值	標準差	變異係數
工作地年薪平均數 WORK_AVG	0.0036	0.01%	3.22%
性別倍率 Gender_Rate	0.5394	1.00%	1.86%
上市櫃公司溢酬 Salary_Company_Prem02	0.1579	0.27%	1.74%
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.0013	0.00%	1.56%
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.0166	0.03%	1.59%
有擔保貸款高額度註記 HighLimit_Secur_Mark_C30	0.0937	0.42%	4.48%
近 6 個月高消費註記 HighSpending_6M_Mark_C30	0.1346	0.51%	3.78%

4.9 專業服務業(C40)子模型多變量分析與模型驗證結果

針對子模型 C40(專業服務業)所應用到的資料集說明請參考 4.4 節，而應用到的資料集及結果輸出名稱列示如下：

資料集說明	資料集名稱
訓練集	TrainData_C40_Output.RData
測試集	TestData_C40_Output.RData
建模樣本	ModelData_C40_Output.RData

4.9.1 模型設定

$$\begin{aligned} \text{收入距離} = & \alpha + \beta_1 \text{ 職稱倍率} + \beta_2 \text{ 教育程度年薪中位數} + \beta_3 \text{ 性別倍率} + \beta_4 \text{ 四師註記} \\ & + \beta_5 \text{ 無擔保貸款總額度} + \beta_6 \text{ 近 6 個月平均每發卡銀行永久額度} \\ & + \beta_7 \text{ 有擔保貸款高額度註記} + \varepsilon \end{aligned}$$

4.9.2 模型結果

以 2,632 筆訓練集資料配適多元線性迴歸，模型係數四捨五入至小數第 8 位，C40 子模型結果如下：

變數名稱	模型係數	T 值	P 值	VIF 值
截距項	-1.06037164	-23.21	<0.01	-
職稱倍率 Jobtitle_Rate	0.17297540	12.09	<0.01	2.25
教育程度年薪中位數 EDU_MID	0.00337548	7.40	<0.01	1.10
性別倍率 Gender_Rate	0.23139878	5.44	<0.01	1.21
四師註記 FourPros_Flag	0.46263754	15.03	<0.01	2.36
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00145651	21.62	<0.01	1.61
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01881871	19.54	<0.01	1.92

變數名稱	模型係數	T 值	P 值	VIF 值
有擔保貸款高額度註記 HighLimit_Secur_Mark_C40	0.15496259	6.91	<0.01	1.29

C40 子模型的 R-squared 值為 0.7456，Adjusted R-squared 值為 0.7449，所有投入變數之模型係數均顯著異於 0。VIF 值均小於 10，變數之間不存在共線性問題。

在業務的應用上，會希望能夠檢視各變數的重要程度，以了解其貢獻性。因此 KPMG 將上述模型進行標準化迴歸，其迴歸係數便能夠揭露其重要程度。結果與變數對應的權重列表如下：

變數名稱	標準化迴歸係數	變數權重
職稱倍率 Jobtitle_Rate	0.1784	15.4%
教育程度年薪中位數 EDU_MID	0.0765	6.6%
性別倍率 Gender_Rate	0.0589	5.1%
四師註記 FourPros_Flag	0.2276	19.7%
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.2701	23.4%
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.2668	23.1%
有擔保貸款高額度註記 HighLimit_Secur_Mark_C40	0.0774	6.7%

根據 C40 子模型的模型結果，投入對應的變數即可預測收入距離，後續透過 4.2 節收入距離公式可推導出預測月收入⁴⁴，結果詳見資料集中的「yhat」欄位。

⁴⁴ 依模型估計得到的預測收入距離(\hat{y}_d)會以 $e^{\hat{y}_d + \log(\text{子模型分類下之總體趨勢薪資}[\text{MONTH_INCOME_ALL_WA}])}$ 轉換為預測月收入

4.9.3 模型效度

下表為子模型分類 C40 的誤差百分比⁴⁵在各區間的樣本佔比以及 MAPE⁴⁶統計結果：

誤差百分比	訓練集	測試集	新增測試集 ⁴⁷	全樣本
-100%~-50%	2.85%	3.34%	6.51%	2.95%
-50%~-20%	22.45%	22.04%	18.01%	22.37%
-20%~-10%	11.66%	10.94%	9.28%	11.52%
-10%~0%	11.02%	10.18%	8.03%	10.85%
0%~10%	10.60%	10.33%	8.31%	10.55%
10%~20%	9.65%	11.25%	9.56%	9.97%
20%~40%	15.27%	15.81%	13.02%	15.38%
40%~50%	5.40%	5.47%	4.99%	5.41%
50%~75%	8.17%	7.60%	10.66%	8.05%
75%~100%	1.71%	1.67%	8.73%	1.70%
>100%	1.22%	1.37%	2.91%	1.25%
樣本整體 MAPE	27.97%	28.19%	37.45%	28.01%
樣本數	2,632	658	722	3,290

下表則是預估月收入減核實月收入誤差之各區間樣本占比：

誤差範圍	訓練集	測試集	新增測試集	全樣本
<-10 萬	3.76%	3.65%	5.54%	3.74%
-10 萬~-5 萬	3.76%	3.65%	2.49%	3.74%
-5 萬~-3 萬	5.17%	5.47%	4.43%	5.23%
-3 萬~-2 萬	6.53%	7.90%	6.65%	6.81%
-2 萬~-1 萬	12.54%	10.79%	9.56%	12.19%
-1 萬~0 萬	16.22%	15.05%	13.16%	15.99%
0 萬~1 萬	23.18%	23.10%	20.64%	23.16%
1 萬~2 萬	18.96%	20.82%	27.29%	19.33%
2 萬~3 萬	3.38%	3.65%	8.17%	3.43%

⁴⁵ 誤差百分比 = (預測月收入-核實月收入) / 核實月收入

⁴⁶ 將「yhat」欄位的預測月收入，與「Monthly_Income」欄位的核實月收入套入 MAPE 的計算公式(詳見 2.2.4 節之模型效度指標)，即可計算出預測月收入與核實月收入間的 MAPE

⁴⁷ 新增測試集為 2023/12/5 依貴行提出需求新增之測試集，擷取樣本範圍是訓練集以外樣本，並排除(1)非信貸客群 (2)近 6M 每發卡銀行永久額度>50 萬之非適用客群 (3)策略預估大於年收>200 萬之非適用客群 (4)貴行授信政策下不予承做之核實年薪 24 萬以下客群

誤差範圍	訓練集	測試集	新增測試集	全樣本
3 萬~5 萬	1.86%	2.28%	1.80%	1.95%
5 萬~10 萬	2.43%	1.98%	0.28%	2.34%
>10 萬	2.20%	1.67%	5.54%	2.10%
平均絕對誤差(元)	25,669.21	25,873.89	38,894.73	25,710.15

由上表結果顯示，C40 子模型大約有 86% 的樣本估計誤差介於 $\pm 50\%$ 之間，而平均絕對誤差⁴⁸約為 25,669 元。每筆樣本的誤差百分比之詳細結果詳見資料集中的「PerErr」欄位，誤差值詳見「Err」欄位。

■ 各收入區間⁴⁹

下表是將全樣本分隔成多個預估年收入區間，並計算在各區間樣本下不同誤差範圍下的樣本占比，以及各區間的樣本數、平均絕對誤差(MAE)、平均絕對誤差百分比(MAPE)。後續的應用策略框架報告書將會再說明 C40 子模型與分量迴歸模型的搭配使用。

誤差範圍 (萬元)	預估年收入區間(萬元)												總計
	0-30	30-40	40-50	50-75	75-100	100-150	150-200	200-250	250-300	300-400	400-500	> 500	
<-10	0.0%	0.0%	0.0%	0.2%	0.3%	0.5%	1.0%	0.7%	0.7%	0.4%	0.0%	0.0%	3.7%
-10~-5	0.0%	0.0%	0.1%	0.6%	0.7%	0.7%	0.4%	0.2%	0.3%	0.6%	0.1%	0.0%	3.7%
-5~-3	0.0%	0.0%	0.2%	2.9%	0.9%	0.3%	0.2%	0.2%	0.1%	0.3%	0.1%	0.0%	5.2%
-3~-2	0.0%	0.0%	0.5%	4.4%	1.1%	0.2%	0.1%	0.2%	0.1%	0.2%	0.0%	0.0%	6.8%
-2~-1	0.0%	0.0%	0.8%	9.1%	1.4%	0.4%	0.2%	0.1%	0.0%	0.3%	0.1%	0.0%	12.2%
-1~0	0.0%	0.0%	1.5%	11.9%	1.9%	0.4%	0.0%	0.1%	0.0%	0.1%	0.1%	0.0%	16.0%
0~1	0.0%	0.0%	3.4%	17.0%	1.7%	0.6%	0.0%	0.2%	0.0%	0.2%	0.1%	0.0%	23.2%
1~2	0.0%	0.0%	2.3%	14.4%	1.7%	0.4%	0.1%	0.1%	0.0%	0.2%	0.1%	0.0%	19.3%
2~3	0.0%	0.0%	0.0%	1.7%	0.9%	0.3%	0.2%	0.1%	0.0%	0.2%	0.2%	0.0%	3.4%
3~5	0.0%	0.0%	0.0%	0.1%	0.3%	0.7%	0.3%	0.1%	0.1%	0.2%	0.2%	0.0%	1.9%
5~10	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.3%	0.4%	0.2%	0.5%	0.4%	0.5%	2.3%
>10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.4%	0.5%	1.0%	2.1%
樣本數	0	0	289	2,048	362	152	91	78	52	111	57	50	3,290
MAE	-	-	10,496	13,015	24,067	43,023	90,298	82,249	101,644	64,421	74,130	167,053	25,710
MAPE	-	-	26.9%	25.7%	27.0%	37.1%	40.1%	35.7%	38.2%	28.0%	31.9%	59.0%	28.0%

⁴⁸ 平均絕對誤差(MAE)即為每一筆樣本預測月收入與核實月收入相減後取絕對值再平均

⁴⁹ 各收入區間是以年收入(ANNINCOME_REVIEW)欄位進行判斷，根據年收入高低劃分成 12 個不同所得級距的區間。

4.9.4 模型變數相關說明

■ 職稱倍率

變數名稱為 Jobtitle_Rate，其變數來源由建模資料集與 3.5 節職類別薪資資料處理產出之內外部職稱與職稱倍率彙總表即 TABLE_JOB，利用下述欄位與邏輯進行資料對應與串接而得：

- 職稱資料匹配年月欄位(ADJ_Mapping_YM)，該欄位為透過資料對照年月(Mapping_YM)，將該欄位拆分為資料年度(YEAR)與資料月份(MONTH)並進行調整後合併，其調整邏輯為考量內外部職稱與職稱倍率彙總表的外部資料來源更新時點為每年度的 5 月底至 6 月初，而公布的資料為去年度的 7 月份，舉例如 2022 年 5 月底發布的外部資料為 2021 年 7 月各行業別下各職稱的薪資，因此 2022 年 7 月至 2023 年 6 月的樣本可套用 2021 年 7 月的職稱薪資統計資料，故在建立內外部職稱與職稱倍率彙總表的「職稱資料匹配年月欄位(ADJ_Mapping_YM)」上，當資料對照年月(Mapping_YM)的月份(MONTH)大於等於 7 月時，職稱資料匹配年月(ADJ_Mapping_YM)為資料年度(YEAR)減去 1 後再與字串"07"合併而得，反之當資料對照年月(Mapping_YM)的月份(MONTH)小於 7 月時，職稱資料匹配年月(ADJ_Mapping_YM)為資料年度(YEAR)減去 2 後再與字串"07"合併而得，最後利用上述產出之職稱資料匹配年月(ADJ_Mapping_YM)符合資料年月(DATA_YM)的邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同年度的資料。
- 主計處行業別代碼(CAREER_CODE)，可由 3.2 節之說明以重新歸類後之行業別與邏輯取得，利用該欄位符合行業別代碼大類(CAREER_CODE)邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同行業別代碼的資料。
- 職稱(JOB_CODE)，可由 3.5 節之說明以重新歸類後之職稱別與邏輯取得，利用該欄位符合行內職稱代碼(JOB_TITLE_CODE_BANK)邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同職稱代碼的資料。

針對上述欄位與邏輯利用迴圈方式同時進行判斷，找出並串接每一筆資料集樣本於內外部職稱與職稱倍率彙總表(TABLE_JOB)中相應的職稱倍率_初始(BENCHMARK_RATE)欄位，最後將該欄位命名為 Jobtitle_Rate。

■ 教育程度年薪中位數

變數名稱為 EDU_MID，變數來源為資料集所彙整的「教育程度(EDUCATION)」欄位。該變數相關處理是以行政院主計總處於 110 年所發布之「工業及服務業受僱員工全年總薪資中位數及分布統計結果」⁵⁰中的表 1-3「工業及服務業受僱員工全年總薪資中位數」統計之年薪中位數(單位：萬元)作為薪資水準進行套用，資料期間共包含了 105~110 年度，KPMG 也將數據彙整於「特徵變數轉換檔」⁵¹的學歷變數總表頁籤中。考量到該份外部資料更新時點為隔年 12 月底，例如：2021 年底發布的外部資料為 2020 年度的薪資統計，因此 2022 年 1 月至 2022 年 12 月的樣本可套用最新發布之 2020 年度的薪資統計，故以下列規則衍生變數：

- APPROVE_DATE 在 2018/1/1~2018/12/31，依據 EDUCATION 所對應的學歷套用 105 年度薪資統計，命名為 EDU_MID_2016
- APPROVE_DATE 在 2019/1/1~2019/12/31，依據 EDUCATION 所對應的學歷套用 106 年度薪資統計，命名為 EDU_MID_2017
- APPROVE_DATE 在 2020/1/1~2020/12/31，依據 EDUCATION 所對應的學歷套用 107 年度薪資統計，命名為 EDU_MID_2018
- APPROVE_DATE 在 2021/1/1~2021/12/31，依據 EDUCATION 所對應的學歷套用 108 年度薪資統計，命名為 EDU_MID_2019
- APPROVE_DATE 在 2022/1/1~2022/12/31，依據 EDUCATION 所對應的學歷套用 109 年度薪資統計，命名為 EDU_MID_2020
- APPROVE_DATE 在 2023/1/1~2023/12/31，依據 EDUCATION 所對應的學歷套用 110 年度薪資統計，命名為 EDU_MID_2021

與資料集進行 JOIN 後再將 EDU_MID_XXXX 欄位進行整併，並命名為 EDU_MID，若 EDUCATION 缺值則同樣給予總計的中位數薪資水準。下圖為部分節錄自附錄 5.6 之特徵變數轉換檔的學歷變數總表：

	A	B	C	D	E	F	G	H
1	EDU_CODE	EDU_CLASS	EDU_MID_2016	EDU_MID_2017	EDU_MID_2018	EDU_MID_2019	EDU_MID_2020	EDU_MID_2021
2		0 研究所	88.4	90.4	94.1	95.2	96	100.2
3		1 大專	51.1	51.8	53.2	53.6	53.7	54.5
4		2 高中	41.2	42.2	43.6	44.2	44.7	44.9
5		3 國中以下	37.3	38	39.4	40.1	40.5	40.7

⁵⁰ 「工業及服務業受僱員工全年總薪資中位數及分布統計結果」請參考附錄 5.5 之 PDF 檔

⁵¹ 「特徵變數轉換檔」的學歷變數總表請參考附錄 5.6 之 EXCEL 檔

■ 性別倍率

變數名稱為 Gender_Rate，變數來源為資料集所彙整的「性別(gender)」欄位。該變數相關處理是以 3.4 節完成之建模資料集⁵²裡的三個欄位「平均月薪_全體(MONTH_INCOME_ALL_WA)」、「平均月薪_男(MONTH_INCOME_M_WA)」以及「平均月薪_女(MONTH_INCOME_F_WA)」進行計算即可取得性別倍率變數，計算邏輯請參考以下：

- 若性別(gender)屬於 1，則將平均月薪_男(MONTH_INCOME_M_WA)除以平均月薪_全體(MONTH_INCOME_ALL_WA)作為性別倍率變數值
- 若性別(gender)屬於 2，則將平均月薪_女(MONTH_INCOME_F_WA)除以平均月薪_全體(MONTH_INCOME_ALL_WA)作為性別倍率變數值

■ 四師註記

變數名稱為 FourPros_Flag，變數來源為資料集所彙整的「職稱(JOB_CODE)」欄位，可由 3.2 節之說明以重新歸類後之職稱別與邏輯取得。該變數相關處理是將職稱(JOB_CODE)欄位符合醫師(代碼為：07_1、08_2、39)、建築師(代碼為：07_2)、會計師(代碼為：40_1)與律師(代碼為：40_2)代碼者賦值為 1，其他則賦值為 0。

■ 無擔保貸款總額度

變數名稱為 VB011_Loan_NonSecur_Limit_NEW_10K_n1，變數來源為 貴行於 2023/9/19 所提供資料表中的「無擔保貸款總額度(VB011_Loan_NonSecur_Limit)」欄位，後續將該欄位重新命名為 VB011_Loan_NonSecur_Limit_NEW，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 VB011_Loan_NonSecur_Limit_NEW 除以 10,000，並且參考同業相關信貸最高可貸額度普遍設定為 500 萬，故設定 500 萬元作為該變數之上限值，即 VB011_Loan_NonSecur_Limit_NEW > 5,000,000，則等於 5,000,000。

■ 近 6 個月平均每發卡銀行永久額度

變數名稱為 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1，變數來源為 貴行於 2023/9/19 所提供資料表中的「近 6 個月平均每發卡銀行永久額度

⁵² 模型 C10~C50 請參考建模資料集-1 名為 ModelTable_Stable_v01

(VK138_Card_Avg_PerBankLimit_6M) 」欄位，後續將該欄位重新命名為 VK138_Card_Avg_PerBankLimit_6M_NEW，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 VK138_Card_Avg_PerBankLimit_6M_NEW 除以 10,000，並且參考同業相關信用卡給額設定，發卡額度若超過 50 萬元一般屬於頂級卡客，於額度申請時須提供較多元完整可認列資產在相當程度以上的證明，故設定 50 萬元作為該變數之上限值，即 VK138_Card_Avg_PerBankLimit_6M_NEW > 500,000，則等於 500,000。

■ 有擔保貸款高額度註記

變數名稱為 HighLimit_Secur_Mark_C40，變數來源為貴行於 2023/9/19 所提供資料表中的「擔保貸款總額度(VB002_Loan_Secur_Limit)」欄位，後續將該欄位重新命名為 VB002_Loan_Secur_Limit_NEW，空值則補 0。該變數額外相關處理是檢視並參考變數來源欄位的每個百分位點之數值並反覆測試模型找到最佳之高額度門檻，當變數值大於等於門檻值則賦值為 1，小於門檻值則賦值為 0，而最終 C40 子模型之高額度門檻訂為 5,000,000 元。

4.9.5 分量迴歸模型結果

以 C40 子模型投入變數為基礎，進行第 40 百分位數(Q40)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.10823007	-20.28	<0.01
職稱倍率 Jobtitle_Rate	0.13846652	7.57	<0.01
教育程度年薪中位數 EDU_MID	0.00297840	5.08	<0.01
性別倍率 Gender_Rate	0.18219235	3.77	<0.01
四師註記 FourPros_Flag	0.37756524	7.96	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00161725	20.87	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.02226872	17.28	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C40	0.16975141	5.95	<0.01

以 C40 子模型投入變數為基礎，進行第 35 百分位數(Q35)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.14033277	-21.93	<0.01
職稱倍率 Jobtitle_Rate	0.13192163	6.69	<0.01
教育程度年薪中位數 EDU_MID	0.00325026	6.71	<0.01
性別倍率 Gender_Rate	0.17514305	3.72	<0.01
四師註記 FourPros_Flag	0.35851301	8.67	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00169411	23.80	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.02070175	15.77	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C40	0.16902862	6.66	<0.01

以 C40 子模型投入變數為基礎，進行第 30 百分位數(Q30)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.18224856	-22.05	<0.01
職稱倍率 Jobtitle_Rate	0.11766514	6.56	<0.01
教育程度年薪中位數 EDU_MID	0.00361593	6.86	<0.01
性別倍率 Gender_Rate	0.17704186	3.59	<0.01
四師註記 FourPros_Flag	0.37198348	8.79	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00176312	23.08	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01946188	14.64	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C40	0.17239938	5.97	<0.01

以 C40 子模型投入變數為基礎，進行第 25 百分位數(Q25)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.23750108	-24.41	<0.01
職稱倍率 Jobtitle_Rate	0.12707049	12.45	<0.01
教育程度年薪中位數 EDU_MID	0.00330493	6.29	<0.01
性別倍率 Gender_Rate	0.20374267	4.47	<0.01
四師註記 FourPros_Flag	0.34568199	9.92	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00184034	23.84	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01825796	15.18	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C40	0.18183733	9.02	<0.01

以 C40 子模型投入變數為基礎，進行第 20 百分位數(Q20)的分量迴歸配適，模型結果如下：

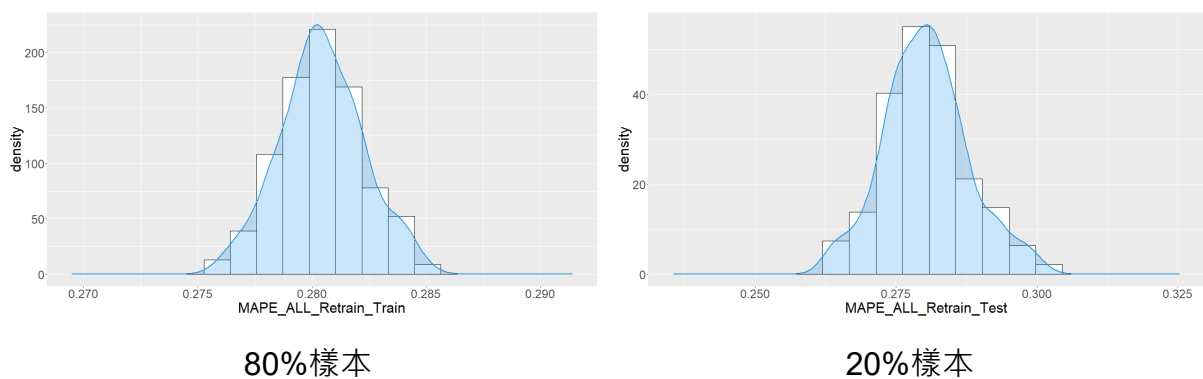
變數名稱	模型係數	T 值	P 值
截距項			<0.01
職稱倍率 Jobtitle_Rate			<0.01
教育程度年薪中位數 EDU_MID			<0.01
性別倍率 Gender_Rate			<0.01
四師註記 FourPros_Flag			<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1			<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1			<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C40			<0.01

4.9.6 拔靴法交叉驗證

■ 模型效度穩定性

根據交叉驗證重新擬合結果，訓練集與測試集樣本的 **MAPE** 分布均集中於平均值，且變異係數小於 30%，顯示 **C40** 子模型的模型效度穩定性佳。

下圖與下表為 200 次 **Bootstrapping** 交叉驗證的試驗結果，每次試驗將樣本根據核實所得區間進行分層抽樣，並隨機劃分出 80% 為訓練集，20% 為測試集，重新訓練 **C40** 子模型後所得到樣本整體 **MAPE** 的分布如下：



樣本範圍	樣本數	MAPE 平均值	MAPE 標準差	MAPE 變異係數
80%樣本	2,632	28.04%	0.18%	0.65%
20%樣本	658	28.03%	0.75%	2.67%

■ 模型係數穩定性

根據交叉驗證重新擬合結果，各模型係數之變異係數均在 30% 以下，顯示 **C40** 子模型的模型係數估計結果具可靠性。

經 200 次 **Bootstrapping** 交叉驗證，每次試驗抽出 80% 的訓練集樣本並重新訓練 **C40** 子模型後，所得到的模型係數分布如下表：

變數名稱	平均值	標準差	變異係數
職稱倍率 Jobtitle_Rate	0.1693	0.87%	5.15%
教育程度年薪中位數 EDU_MID	0.0036	0.02%	5.77%
性別倍率 Gender_Rate	0.2351	1.93%	8.20%
四師註記 FourPros_Flag	0.4776	1.69%	3.53%

變數名稱	平均值	標準差	變異係數
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.0014	0.00%	2.71%
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.0192	0.05%	2.79%
有擔保貸款高額度註記 HighLimit_Secur_Mark_C40	0.1567	1.05%	6.70%

4.10 資通金融能源業(C50)子模型多變量分析與模型驗證結果

針對子模型 C50(資通金融能源業)所應用到的資料集說明請參考 4.4 節，而應用到的資料集及結果輸出名稱列示如下：

資料集說明	資料集名稱
訓練集	TrainData_C50_Output.RData
測試集	TestData_C50_Output.RData
建模樣本	ModelData_C50_Output.RData

4.10.1 模型設定

收入距離 = $\alpha + \beta_1$ 職稱倍率 + β_2 推估總年資 + β_3 上市櫃公司溢酬 + β_4 工作地年薪平均數 + β_5 性別倍率 + β_6 無擔保貸款總額度 + β_7 近 6 個月平均每發卡銀行永久額度 + β_8 有擔保貸款高額度註記 + β_9 近 12 個月高消費註記 + β_{10} 外商金融機構註記 + ε

4.10.2 模型結果

以 2,979 筆訓練集資料配適多元線性迴歸，模型係數四捨五入至小數第 8 位，C50 子模型結果如下：

變數名稱	模型係數	T 值	P 值	VIF 值
截距項	-2.40896110	-22.23	<0.01	-
職稱倍率 Jobtitle_Rate	0.15054300	21.12	<0.01	1.47
推估總年資 SENIORITY_YEAR_TOT	0.01314424	19.09	<0.01	1.36
上市櫃公司溢酬 Salary_Company_Prem02	0.27903716	7.88	<0.01	1.05
工作地年薪平均數 WORK_AVG	0.00261208	5.38	<0.01	1.04
性別倍率 Gender_Rate	0.91327980	9.43	<0.01	1.03

變數名稱	模型係數	T 值	P 值	VIF 值
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00107774	19.29	<0.01	1.23
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01765354	22.02	<0.01	1.71
有擔保貸款高額度註記 HighLimit_Secur_Mark_C50	0.11598173	7.65	<0.01	1.18
近 12 個月高消費註記 HighSpending_12M_Mark_C50	0.15756684	9.37	<0.01	1.28
外商金融機構註記 GlobalCompany_Mark	0.08990237	4.05	<0.01	1.06

C50 子模型的 R-squared 值為 0.7021，Adjusted R-squared 值為 0.7011，所有投入變數之模型係數均顯著異於 0。VIF 值均小於 10，變數之間不存在共線性問題。

在業務的應用上，會希望能夠檢視各變數的重要程度，以了解其貢獻性。因此 KPMG 將上述模型進行標準化迴歸，其迴歸係數便能夠揭露其重要程度。結果與變數對應的權重列表如下：

變數名稱	標準化迴歸係數	變數權重
職稱倍率 Jobtitle_Rate	0.2563	17.7%
推估總年資 SENIORITY_YEAR_TOT	0.2232	15.4%
上市櫃公司溢酬 Salary_Company_Prem02	0.0810	5.6%
工作地年薪平均數 WORK_AVG	0.0549	3.8%
性別倍率 Gender_Rate	0.0960	6.6%
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.2140	14.8%
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.2888	20.0%
有擔保貸款高額度註記 HighLimit_Secur_Mark_C50	0.0831	5.8%

變數名稱	標準化迴歸係數	變數權重
近 12 個月高消費註記 HighSpending_12M_Mark_C50	0.1061	7.3%
外商金融機構註記 GlobalCompany_Mark	0.0419	2.9%

根據 C50 子模型的模型結果，投入對應的變數即可預測收入距離，後續透過 4.2 節收入距離公式可推導出預測月收入⁵³，結果詳見資料集中的「yhat」欄位。

4.10.3 模型效度

下表為子模型分類 C50 的誤差百分比⁵⁴在各區間的樣本佔比以及 MAPE⁵⁵統計結果：

誤差百分比	訓練集	測試集	新增測試集 ⁵⁶	全樣本
-100%~-50%	1.88%	2.68%	3.85%	2.04%
-50%~-20%	18.40%	17.45%	16.78%	18.21%
-20%~-10%	11.38%	13.15%	12.93%	11.73%
-10%~0%	16.35%	13.69%	13.62%	15.82%
0%~10%	13.33%	13.96%	13.62%	13.45%
10%~20%	11.61%	12.35%	11.69%	11.76%
20%~40%	16.05%	14.23%	13.34%	15.68%
40%~50%	4.13%	4.83%	4.81%	4.27%
50%~75%	4.70%	5.10%	5.64%	4.78%
75%~100%	1.11%	1.21%	2.20%	1.13%
>100%	1.07%	1.34%	1.51%	1.13%
樣本整體 MAPE	23.37%	24.37%	26.17%	23.57%
樣本數	2,979	745	727	3,724

⁵³ 依模型估計得到的預測收入距離(\hat{y}_a)會以 $e^{\hat{y}_a + \log(\text{子模型分類下之總體趨勢薪資}[\text{MONTH_INCOME_ALL_WA}])}$ 轉換為預測月收入

⁵⁴ 誤差百分比 = (預測月收入-核實月收入) / 核實月收入

⁵⁵ 將「yhat」欄位的預測月收入，與「Monthly_Income」欄位的核實月收入套入 MAPE 的計算公式(詳見 2.2.4 節之模型效度指標)，即可計算出預測月收入與核實月收入間的 MAPE

⁵⁶ 新增測試集為 2023/12/5 依貴行提出需求新增之測試集，擷取樣本範圍是訓練集以外樣本，並排除(1)非信貸客群 (2)無擔總額度>500 萬之非適用客群 (3)近 6M 每發卡銀行永久額度>50 萬之非適用客群 (3)策略預估大於年收>200 萬之非適用客群 (4)貴行授信政策下不予承做之核實年薪 24 萬以下客群

下表則是預估月收入減核實月收入誤差之各區間樣本占比：

誤差範圍	訓練集	測試集	新增測試集	全樣本
<-10 萬	1.88%	1.74%	2.48%	1.85%
-10 萬~-5 萬	4.67%	5.64%	5.23%	4.86%
-5 萬~-3 萬	6.81%	6.17%	5.78%	6.69%
-3 萬~-2 萬	6.68%	5.23%	5.09%	6.39%
-2 萬~-1 萬	9.67%	11.54%	11.83%	10.04%
-1 萬~0 萬	18.29%	16.64%	16.78%	17.96%
0 萬~1 萬	23.87%	24.16%	24.62%	23.93%
1 萬~2 萬	15.61%	13.96%	16.92%	15.28%
2 萬~3 萬	5.98%	6.31%	6.74%	6.04%
3 萬~5 萬	3.59%	4.70%	3.85%	3.81%
5 萬~10 萬	1.85%	2.42%	0.69%	1.96%
>10 萬	1.11%	1.48%	2.48%	1.18%
平均絕對誤差(元)	21,589.39	23,344.34	24,701.43	21,940.47

由上表結果顯示，C50 子模型大約有 91% 的樣本估計誤差介於 $\pm 50\%$ 之間，而平均絕對誤差⁵⁷約為 21,589 元。每筆樣本的誤差百分比之詳細結果詳見資料集中的「PerErr」欄位，誤差值詳見「Err」欄位。

■ 各收入區間⁵⁸

下表是將全樣本分隔成多個預估年收入區間，並計算在各區間樣本下不同誤差範圍下的樣本占比，以及各區間的樣本數、平均絕對誤差(MAE)、平均絕對誤差百分比(MAPE)。後續的應用策略框架報告書將會再說明 C50 子模型與分量迴歸模型的搭配使用。

誤差範圍 (萬元)	預估年收入區間(萬元)												總計
	0-30	30-40	40-50	50-75	75-100	100-150	150-200	200-250	250-300	300-400	400-500	> 500	
<-10	0.0%	0.0%	0.0%	0.1%	0.1%	0.8%	0.5%	0.2%	0.2%	0.1%	0.0%	0.0%	1.9%
-10~-5	0.0%	0.0%	0.0%	0.4%	1.4%	1.6%	0.9%	0.2%	0.2%	0.1%	0.0%	0.0%	4.9%
-5~-3	0.0%	0.0%	0.1%	1.2%	2.5%	2.0%	0.6%	0.2%	0.1%	0.0%	0.0%	0.0%	6.7%
-3~-2	0.0%	0.0%	0.0%	2.0%	1.5%	1.9%	0.6%	0.2%	0.1%	0.2%	0.0%	0.0%	6.4%
-2~-1	0.0%	0.0%	0.2%	4.0%	3.0%	2.1%	0.4%	0.2%	0.1%	0.0%	0.0%	0.0%	10.0%

⁵⁷ 平均絕對誤差(MAE)即為每一筆樣本預測月收入與核實月收入相減後取絕對值再平均

⁵⁸ 各收入區間是以年收入(ANNINCOME_REVIEW)欄位進行判斷，根據年收入高低劃分成 12 個不同所得級距的區間。

-1~0	0.0%	0.1%	1.4%	8.5%	3.9%	3.2%	0.5%	0.2%	0.1%	0.1%	0.0%	0.0%	18.0%
0~1	0.0%	0.1%	2.4%	13.1%	4.7%	2.4%	0.8%	0.3%	0.1%	0.1%	0.0%	0.0%	23.9%
1~2	0.0%	0.0%	0.9%	6.7%	4.5%	2.3%	0.6%	0.1%	0.1%	0.1%	0.0%	0.1%	15.3%
2~3	0.0%	0.0%	0.0%	1.2%	2.2%	1.9%	0.4%	0.2%	0.1%	0.1%	0.0%	0.0%	6.0%
3~5	0.0%	0.0%	0.0%	0.0%	0.6%	1.6%	0.8%	0.4%	0.1%	0.2%	0.1%	0.0%	3.8%
5~10	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.5%	0.6%	0.3%	0.3%	0.1%	0.0%	2.0%
>10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.6%	0.2%	0.2%	1.2%
樣本數	0	10	185	1,385	904	743	238	102	62	65	19	11	3,724
MAE	-	4,308	6,706	10,987	19,220	28,371	42,558	46,931	53,047	73,995	100,689	165,755	21,940
MAPE	-	12.9%	19.2%	20.8%	24.3%	24.2%	26.7%	28.4%	28.1%	43.0%	42.8%	68.0%	23.6%

4.10.4 模型變數相關說明

■ 職稱倍率

變數名稱為 Jobtitle_Rate，其變數來源由建模資料集與 3.5 節職類別薪資資料處理產出之內外部職稱與職稱倍率彙總表即 TABLE_JOB，利用下述欄位與邏輯進行資料對應與串接而得：

- 職稱資料匹配年月欄位(ADJ_Mapping_YM)，該欄位為透過資料對照年月(Mapping_YM)，將該欄位拆分為資料年度(YEAR)與資料月份(MONTH)並進行調整後合併，其調整邏輯為考量內外部職稱與職稱倍率彙總表的外部資料來源更新時點為每年度的 5 月底至 6 月初，而公布的資料為去年度的 7 月份，舉例如 2022 年 5 月底發布的外部資料為 2021 年 7 月各行業別下各職稱的薪資，因此 2022 年 7 月至 2023 年 6 月的樣本可套用 2021 年 7 月的職稱薪資統計資料，故在建立內外部職稱與職稱倍率彙總表的「職稱資料匹配年月欄位(ADJ_Mapping_YM)」上，當資料對照年月(Mapping_YM)的月份(MONTH)大於等於 7 月時，職稱資料匹配年月(ADJ_Mapping_YM)為資料年度(YEAR)減去 1 後再與字串"07"合併而得，反之當資料對照年月(Mapping_YM)的月份(MONTH)小於 7 月時，職稱資料匹配年月(ADJ_Mapping_YM)為資料年度(YEAR)減去 2 後再與字串"07"合併而得，最後利用上述產出之職稱資料匹配年月(ADJ_Mapping_YM)符合資料年月(DATA_YM)的邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同年度的資料。
- 主計處行業別代碼(CAREER_CODE)，可由 3.2 節之說明以重新歸類後之行業別與邏輯取得，利用該欄位符合行業別代碼大類(CAREER_CODE)邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同行業別代碼的資料。

- 職稱(JOB_CODE)·可由 3.5 節之說明以重新歸類後之職稱別與邏輯取得·利用該欄位符合行內職稱代碼(JOB_TITLE_CODE_BANK) 邏輯於內外部職稱與職稱倍率彙總表(TABLE_JOB)找出相同職稱代碼的資料。

針對上述欄位與邏輯利用迴圈方式同時進行判斷·找出並串接每一筆資料集樣本於內外部職稱與職稱倍率彙總表(TABLE_JOB)中相應的職稱倍率_初始(BENCHMARK_RATE)欄位·最後將該欄位命名為 Jobtitle_Rate。

■ 推估總年資

變數名稱為 SENIORITY_Mark_TOT·變數來源為建模資料集所彙整的「年齡(AGE)」與「教育程度(EDUCATION)」欄位·該變數相關處理邏輯參考如下：

- 若教育程度(EDUCATION)屬於 0(研究所以上)者·則將年齡(AGE)減去 24 作為推估總年資(SENIORITY_YEAR_TOT)變數值
- 若教育程度(EDUCATION)屬於 1(大學/大專)者·則將年齡(AGE)減去 22 作為推估總年資(SENIORITY_YEAR_TOT)變數值
- 若教育程度(EDUCATION)屬於 2(高中)者·則將年齡(AGE)減去 18 作為推估總年資(SENIORITY_YEAR_TOT)變數值
- 若教育程度(EDUCATION)屬於 3(國中以下)者·則將年齡(AGE)減去 15 作為推估總年資(SENIORITY_YEAR_TOT)變數值

實際以上述邏輯方式推估·推估總年資會有 1~2 年誤差·舉例如教育程度(EDUCATION)屬於 0(研究所以上)·但年齡(AGE)為 22 者·如遇該情況(於 C50 子模型共 10 筆)則以 0 值做取代。

■ 工作地年薪平均數

變數名稱為 WORK_AVG·變數來源為資料集所彙整的「公司郵遞區號(POST_CODE_COMPANY)」欄位。該變數相關處理是以行政院主計總處於 109 發布之「各縣市工業及服務業企業僱用員工薪資統計說明」⁵⁹中表 1·以及 110 年所發布之「工業及服務業受僱員工全年總薪資中位數及分布統計結果」⁶⁰中表 2-1「各縣市工業及服務業企業僱用員工全年總薪資」統計之年薪資平均數(單位：萬元)作為薪資水準進

⁵⁹「各縣市工業及服務業企業僱用員工薪資統計說明」請參考附錄 5.4 之 PDF 檔

⁶⁰「工業及服務業受僱員工全年總薪資中位數及分布統計結果」請參考附錄 5.5 之 PDF 檔

行套用，資料期間共包含了 107~110 年度，KPMG 也將其數據彙整於「特徵變數轉換檔」⁶¹的地理區域變數總表頁籤中。

考量到該份外部資料更新時點為隔年 12 月底，例如：2021 年底發布的外部資料為 2020 年度的薪資統計，因此 2022 年 1 月至 2022 年 12 月的樣本可套用最新發布之 2020 年度的薪資統計，故以下列規則衍生變數：

- APPROVE_DATE 在 2018/01/31~2020/12/31，依據 POST_CODE_COMPANY 所在縣市套用 107 年度薪資統計，命名為 WORK_AVG_2018
- APPROVE_DATE 在 2021/1/1~2021/12/31，依據 POST_CODE_COMPANY 所在縣市套用 108 年度薪資統計，命名為 WORK_AVG_2019
- APPROVE_DATE 在 2022/1/1~2022/12/31，依據 POST_CODE_COMPANY 所在縣市套用 109 年度薪資統計，命名為 WORK_AVG_2020
- APPROVE_DATE 在 2023/1/1~2023/12/31，依據 POST_CODE_COMPANY 所在縣市套用 110 年度薪資統計，命名為 WORK_AVG_2021

惟因部分外島地區未有相關薪資統計，因此外島部分如金門縣及連江縣採用總計的薪資水準進行取代。與資料集進行 JOIN 後再將 WORK_AVG_XXXX 欄位進行整併，並命名為 WORK_AVG，若 POST_CODE_COMPANY 缺值則同樣給予總計的平均薪資水準。下圖為部分節錄自附錄 5.6 之特徵變數轉換檔的地理區域變數總表：

61

「特徵變數轉換檔」的地理區域變數總表請參考附錄 5.6 之 EXCEL 檔

	A	B	N	O	P	Q
1	ZIP_CODE	ZIP_DESC	WORK_AVG_2018	WORK_AVG_2019	WORK_AVG_2020	WORK_AVG_2021
2	100	臺北市中正區	82.1	84.5	86.3	88.4
3	103	臺北市大同區	82.1	84.5	86.3	88.4
4	104	臺北市中山區	82.1	84.5	86.3	88.4
5	105	臺北市松山區	82.1	84.5	86.3	88.4
6	106	臺北市大安區	82.1	84.5	86.3	88.4
7	108	臺北市萬華區	82.1	84.5	86.3	88.4
8	110	臺北市信義區	82.1	84.5	86.3	88.4
9	111	臺北市士林區	82.1	84.5	86.3	88.4
10	112	臺北市北投區	82.1	84.5	86.3	88.4
11	114	臺北市內湖區	82.1	84.5	86.3	88.4
12	115	臺北市南港區	82.1	84.5	86.3	88.4
13	116	臺北市文山區	82.1	84.5	86.3	88.4
14	200	基隆市仁愛區	60.9	62	62.1	62.9
15	201	基隆市信義區	60.9	62	62.1	62.9
16	202	基隆市中正區	60.9	62	62.1	62.9
17	203	基隆市中山區	60.9	62	62.1	62.9
18	204	基隆市安樂區	60.9	62	62.1	62.9
19	205	基隆市暖暖區	60.9	62	62.1	62.9
20	206	基隆市七堵區	60.9	62	62.1	62.9
21	207	新北市萬里區	61.6	63.3	63.7	65.2
22	208	新北市金山區	61.6	63.3	63.7	65.2
23	220	新北市板橋區	61.6	63.3	63.7	65.2
24	221	新北市汐止區	61.6	63.3	63.7	65.2
25	222	新北市深坑區	61.6	63.3	63.7	65.2
26	223	新北市石碇區	61.6	63.3	63.7	65.2
27	224	新北市瑞芳區	61.6	63.3	63.7	65.2
28	226	新北市平溪區	61.6	63.3	63.7	65.2
29	227	新北市雙溪區	61.6	63.3	63.7	65.2
30	228	新北市貢寮區	61.6	63.3	63.7	65.2

■ 性別倍率

變數名稱為 **Gender_Rate**，變數來源為資料集所彙整的「性別(gender)」欄位。該變數相關處理是以 3.4 節完成之建模資料集⁶²裡的三個欄位「平均月薪_全體(MONTH_INCOME_ALL_WA)」、「平均月薪_男(MONTH_INCOME_M_WA)」以及「平均月薪_女(MONTH_INCOME_F_WA)」進行計算即可取得性別倍率變數，計算邏輯請參考以下：

- 若性別(gender)屬於 1，則將平均月薪_男(MONTH_INCOME_M_WA)除以平均月薪_全體(MONTH_INCOME_ALL_WA)作為性別倍率變數值
- 若性別(gender)屬於 2，則將平均月薪_女(MONTH_INCOME_F_WA)除以平均月薪_全體(MONTH_INCOME_ALL_WA)作為性別倍率變數值

⁶² 模型 C10~C50 請參考建模資料集-1 名為 ModelTable_Stable_v01

■ 上市櫃公司溢酬

變數名稱為 `Salary_Company_Prem02`，變數來源為資料集所彙整的「公司名稱(COMPANY)」、「主計處薪資水準(MONTH_INCOME_CAREERCODE)」欄位。

首先，須將資料處理篇所整理的「上市櫃公司名單列表」與資料集進行 JOIN，若資料集公司名稱(COMPANY)欄位中含有「上市櫃公司名單列表」所列公司名稱關鍵字串者，則賦予該樣本對應的公司代碼(Symbol)，惟派遣、工讀生等職位之敘薪與其他正職員工相異甚大，其薪資水準不宜視同上市櫃公司之正職員工，故當 `OC_CODE_REVIEW` 為 28, 29, 30, 31, 32, 33, 34, 35, 37, 55, 63, 67, 68 時，公司代碼(Symbol)欄位為空。

其次，再依據公司代碼(Symbol)以及資料對照年月(Mapping_YM)，與「上市櫃員工薪資對照表」進行匹配，即當公司代碼(Symbol)相同且資料對照年月(Mapping_YM)介於套用起點與套用終點時，給予「非擔任主管職務之全時員工資訊_員工薪資_平均數_仟元每人(Salary_Company)」欄位的值作為其所屬公司的薪資水準，並將其乘以 1,000 再除以 12 轉換為平均月收入(單位：元)。

最後，設定一個新的 `Salary_Company_Prem` 欄位，當該客戶為上市櫃公司正職員工，即 `Salary_Company` 非空值時，則賦值為 `Salary_Company` 當作該公司薪資水準；反之，當該客戶非屬上市櫃公司正職員工時，則賦值為該行業的主計處薪資水準(MONTH_INCOME_CAREERCODE)作為其所屬產業的薪資水準。最終再將 `Salary_Company_Prem` 除以主計處薪資水準(MONTH_INCOME_CAREERCODE)，即可得到上市櫃公司溢酬。

■ 無擔保貸款總額度

變數名稱為 `VB011_Loan_NonSecur_Limit_NEW_10K_n1`，變數來源為貴行於 2023/9/19 所提供資料表中的「無擔保貸款總額度(VB011_Loan_NonSecur_Limit)」欄位，後續將該欄位重新命名為 `VB011_Loan_NonSecur_Limit_NEW`，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 `VB011_Loan_NonSecur_Limit_NEW` 除以 10,000，並且參考同業相關信貸最高可貸額度普遍設定為 500 萬，故設定 500 萬元作為該變數之上限值，即 `VB011_Loan_NonSecur_Limit_NEW > 5,000,000`，則等於 5,000,000。

■ 近 6 個月平均每發卡銀行永久額度

變數名稱為 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1，變數來源為 貴行於 2023/9/19 所提供資料表中的「近 6 個月平均每發卡銀行永久額度 (VK138_Card_Avg_PerBankLimit_6M)」欄位，後續將該欄位重新命名為 VK138_Card_Avg_PerBankLimit_6M_NEW，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 VK138_Card_Avg_PerBankLimit_6M_NEW 除以 10,000，並且參考同業相關信用卡給額設定，發卡額度若超過 50 萬元一般屬於頂級卡客，於額度申請時須提供較多元完整可認列資產在相當程度以上的證明，故設定 50 萬元作為該變數之上限值，即 $VK138_Card_Avg_PerBankLimit_6M_NEW > 500,000$ ，則等於 500,000。

■ 有擔保貸款高額度註記

變數名稱為 HighLimit_Secur_Mark_C50，變數來源為 貴行於 2023/9/19 所提供資料表中的「擔保貸款總額度(VB002_Loan_Secur_Limit)」欄位，後續將該欄位重新命名為 VB002_Loan_Secur_Limit_NEW，空值則補 0。該變數額外相關處理是檢視並參考變數來源欄位的每個百分位點之數值並反覆測試模型找到最佳之高額度門檻，當變數值大於等於門檻值則賦值為 1，小於門檻值則賦值為 0，而最終 C50 子模型之高額度門檻訂為 5,000,000 元。

■ 近 12 個月高消費註記

變數名稱為 HighSpending_12M_Mark_C20，變數來源為 貴行於 2023/9/19 所提供資料表中的「近 6 個月最高消費金額(VK060_Card_Max_Spending_12M)」欄位，後續將該欄位重新命名為 VK060_Card_Max_Spending_12M_NEW_10K，空值則補 0。該變數額外相關處理是檢視並參考變數來源欄位的每個百分位點之數值並反覆測試模型找到最佳之高消費門檻，當變數值大於等於門檻值則賦值為 1，小於門檻值則賦值為 0，而最終 C50 子模型之高消費門檻訂為 150,000 元。

■ 外商金融機構註記

變數名稱為 GlobalCompany_Mark。該變數的處理方式是從建模資料集中子模型分類代碼(MODEL_CODE)為 C50 之樣本的公司名稱(COMPANY)欄位中，搜尋含有外商金融機構名稱的關鍵字串，此處所指外商金融機構主要是外商來台設立或投資的金融機構，

範圍包含：銀行⁶³、壽險⁶⁴、產險⁶⁵、再保險⁶⁶、證券商⁶⁷、投信商⁶⁸、投顧商與國際發卡組織，並根據欄位中實際輸入的公司名稱決定合適的關鍵字長度，決定的原則基本上有以下兩點：(1)避免名稱混淆，例如：「新加坡商大華銀行」若以「大華」作為搜尋的關鍵字，則本國投顧商「大華國際投顧」也可能會符合搜尋條件而導致混淆，故以「大華銀」作為關鍵字辨識，(2)取最大的交集，例如：「新加坡商星展銀行」在本欄位中的字串有「星展銀行」、「星展台灣商業銀行股份有限公司」、「星展(台灣)商業銀行股份有限公司」、「新加坡商星展銀行股份有限公司」等 7 種，故以最低可辨識的字串長度「星展」作為關鍵字辨識。當無法在這兩個原則下以唯一的關鍵字辨識時，則考慮以多個關鍵字進行辨識。最終彙整的外商金融機構關鍵字字串如下表，當公司名稱(COMPANY)欄位包含下表字串時則賦值為 1，其餘不包含下表字串或欄位為空的則賦值為 0。

變數名稱	字串列表	納入或排除
GlobalCompany_Mark	"瑞穗", "美國銀", "盤谷", "菲律賓首都", "紐約梅隆", "大華銀", "道富", "法國興業", "德意志", "東亞", "摩根", "星展", "法國巴黎", "渣打", "華僑", "東方匯理", "瑞士銀", "瑞士商業", "安智", "富國銀", "三菱日聯", "三菱東京日聯", "三井住友", "花旗", "匯豐", "西班牙對外", "澳盛", "法國外貿", "印尼人民", "中國銀", "交通銀", "中國建設", "友邦", "安達", "安聯", "保誠", "康健", "保德信", "科法斯", "美國國際", "裕利安宜", "美國再保", "科隆再保", "麥格理", "美林", "高盛", "瑞士信貸", "野村", "瑞銀", "大和", "安本標準", "宏利", "貝萊德", "施羅德", "柏瑞", "富達證券", "富蘭克林", "景順", "路博邁", "德銀", "鋒裕匯理", "聯博", "瀚亞", "安睿", "美盛", "亨達", "霸菱", "百達", "核聚", "駿利亨德森", "品浩太平洋", "瑞聯", "威士", "萬事達", "吉士美", "美國運通", "銀聯", "環匯"	納入

⁶³ 外商銀行的名單參考以下網址：

<https://www.ba.org.tw/Related/RelatedSiteList?type=RelateForeignbanks&AspxAutoDetectCookieSupport=1>

⁶⁴ 外商壽險公司的名單參考以下網址：<https://www.ib.gov.tw/ch/home.jsp?id=181&parentpath=0,9,179>

⁶⁵ 外商產險公司的名單參考以下網址：<https://www.ib.gov.tw/ch/home.jsp?id=180&parentpath=0,9,179>

⁶⁶ 外商再保險公司的名單參考以下網址：<https://www.ib.gov.tw/ch/home.jsp?id=206&parentpath=0,9,179>

⁶⁷ 證券商的名單是參考以下網址，再從名單中查詢並篩選出具有外商背景的證券商：

<https://www.twse.com.tw/zh/brokerService/brokerServiceAudit>

⁶⁸ 投信商與投顧商皆參考以下網址，再從名單中查詢並篩選出具有外商背景的投信商與投顧商：

<https://www.sitca.org.tw/ROC/MemData/MD2001N.aspx?PGMID=AS0701>

4.10.5 分量迴歸模型結果

以 C50 子模型投入變數為基礎，進行第 40 百分位數(Q40)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-2.33961709	-19.37	<0.01
職稱倍率 Jobtitle_Rate	0.14934134	18.14	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.01330836	16.53	<0.01
上市櫃公司溢酬 Salary_Company_Prem02	0.32171791	7.85	<0.01
工作地年薪平均數 WORK_AVG	0.00201792	3.62	<0.01
性別倍率 Gender_Rate	0.73281293	6.70	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00118385	18.78	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01898204	19.00	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C50	0.11740762	6.39	<0.01
近 12 個月高消費註記 HighSpending_12M_Mark_C50	0.14260749	6.33	<0.01
外商金融機構註記 GlobalCompany_Mark	0.08768595	3.04	<0.01

以 C50 子模型投入變數為基礎，進行第 35 百分位數(Q35)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-2.35644695	-20.22	<0.01
職稱倍率 Jobtitle_Rate	0.15004693	15.18	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.01248954	15.99	<0.01
上市櫃公司溢酬 Salary_Company_Prem02	0.31798877	8.09	<0.01
工作地年薪平均數 WORK_AVG	0.00209176	3.94	<0.01
性別倍率	0.72493666	6.81	<0.01

變數名稱	模型係數	T 值	P 值
Gender_Rate			
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00125633	18.40	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01877194	17.87	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C50	0.11856256	6.50	<0.01
近 12 個月高消費註記 HighSpending_12M_Mark_C50	0.12372078	5.37	<0.01
外商金融機構註記 GlobalCompany_Mark	0.07680125	2.46	0.014

以 C50 子模型投入變數為基礎，進行第 30 百分位數(Q30)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-2.25287483	-19.05	<0.01
職稱倍率 Jobtitle_Rate	0.14720248	11.12	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.01289539	16.43	<0.01
上市櫃公司溢酬 Salary_Company_Prem02	0.31818229	7.80	<0.01
工作地年薪平均數 WORK_AVG	0.00183017	3.45	<0.01
性別倍率 Gender_Rate	0.61648325	5.77	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00125470	16.43	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01768313	16.59	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C50	0.11781261	6.04	<0.01
近 12 個月高消費註記 HighSpending_12M_Mark_C50	0.12685980	5.78	<0.01
外商金融機構註記 GlobalCompany_Mark	0.06246198	2.11	0.035

以 C50 子模型投入變數為基礎，進行第 25 百分位數(Q25)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-2.33428708	-21.93	<0.01
職稱倍率 Jobtitle_Rate	0.13274229	10.21	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.01238305	16.98	<0.01
上市櫃公司溢酬 Salary_Company_Prem02	0.29738868	7.85	<0.01
工作地年薪平均數 WORK_AVG	0.00202674	4.17	<0.01
性別倍率 Gender_Rate	0.69312655	7.32	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00126214	18.22	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01792698	18.48	<0.01
有擔保貸款高額度註記 HighLimit_Secur_Mark_C50	0.12011180	6.43	<0.01
近 12 個月高消費註記 HighSpending_12M_Mark_C50	0.12285771	6.02	<0.01
外商金融機構註記 GlobalCompany_Mark	0.07077555	4.16	<0.01

以 C50 子模型投入變數為基礎，進行第 20 百分位數(Q20)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-2.27396902	-18.43	<0.01
職稱倍率 Jobtitle_Rate	0.12715391	13.03	<0.01
推估總年資 SENIORITY_YEAR_TOT	0.01211390	13.57	<0.01
上市櫃公司溢酬 Salary_Company_Prem02	0.31277406	6.97	<0.01
工作地年薪平均數 WORK_AVG	0.00220277	3.83	<0.01
性別倍率 Gender_Rate	0.59603228	5.37	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.00128761	16.39	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.01651919	14.02	<0.01

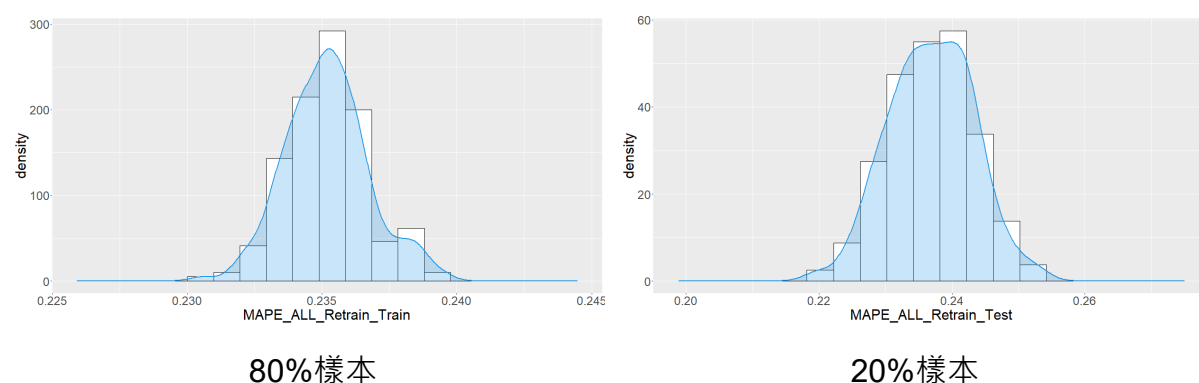
變數名稱	模型係數	T 值	P 值
有擔保貸款高額度註記 HighLimit_Secur_Mark_C50	0.10999391	5.60	<0.01
近 12 個月高消費註記 HighSpending_12M_Mark_C50	0.12817924	5.62	<0.01
外商金融機構註記 GlobalCompany_Mark	0.08291298	4.86	<0.01

4.10.6 拔靴法交叉驗證

■ 模型效度穩定性

根據交叉驗證重新擬合結果，訓練集與測試集樣本的 MAPE 分布均集中於平均值，且變異係數小於 30%，顯示 C50 子模型的模型效度穩定性佳。

下圖與下表為 200 次 Bootstrapping 交叉驗證的試驗結果，每次試驗將樣本根據核實所得區間進行分層抽樣，並隨機劃分出 80% 為訓練集，20% 為測試集，重新訓練 C50 子模型後所得到樣本整體 MAPE 的分布如下：



樣本範圍	樣本數	MAPE 平均值	MAPE 標準差	MAPE 變異係數
80%樣本	2,979	23.52%	0.15%	0.66%
20%樣本	745	23.69%	0.63%	2.68%

■ 模型係數穩定性

根據交叉驗證重新擬合結果，各模型係數之變異係數均在 30% 以下，顯示 C50 子模型的模型係數估計結果具可靠性。

經 200 次 Bootstrapping 交叉驗證，每次試驗抽出 80% 的訓練集樣本並重新訓練 C50 子模型後，所得到的模型係數分布如下表：

變數名稱	平均值	標準差	變異係數
職稱倍率 Jobtitle_Rate	0.1466	0.39%	2.66%
推估總年資 SENIORITY_YEAR_TOT	0.0130	0.03%	2.42%
上市櫃公司溢酬 Salary_Company_Prem02	0.2908	1.55%	5.32%
工作地年薪平均數 WORK_AVG	0.0027	0.02%	8.16%
性別倍率 Gender_Rate	0.8983	4.42%	4.92%
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_10K_n1	0.0011	0.00%	2.95%
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_10K_n1	0.0176	0.04%	2.17%
有擔保貸款高額度註記 HighLimit_Secur_Mark_C50	0.1128	0.73%	6.49%
近 12 個月高消費註記 HighSpending_12M_Mark_C50	0.1577	0.92%	5.84%
外商金融機構註記 GlobalCompany_Mark	0.1013	1.15%	11.33%

4.11 軍公教警類(C60)子模型多變量分析與模型驗證結果

針對子模型分類 C60(軍公教警類)所應用到的資料集說明請參考 4.4 節，而應用到的資料集及結果輸出名稱列示如下：

資料集說明	資料集名稱
訓練集	TrainingData_C60_Output.RData
測試集	TestingData_C60_Output.RData
建模樣本	ModelData_C60_Output.RData

4.11.1 模型設定

$$\begin{aligned}
 \text{收入距離} = & \alpha + \beta_1 \text{離島服務註記} + \beta_2 \log(\text{教育程度年薪中位數}) + \beta_3 \text{年齡} \\
 & + \beta_4 \text{行業別虛擬變數(軍人)} + \beta_5 \text{行業別虛擬變數(警察)} \\
 & + \beta_6 \text{行業別虛擬變數(教師)} + \beta_7 \text{無擔保貸款總額度} \\
 & + \beta_8 \text{近 12 個月平均每發卡銀行永久額度} + \beta_9 \text{信用卡一年期總消費金額} \\
 & + \beta_{10} \text{有擔保貸款高額度註記} + \varepsilon
 \end{aligned}$$

4.11.2 模型結果

以 2,296 筆訓練集資料配適多元線性迴歸，模型係數四捨五入至小數第 8 位，C60 子模型結果如下：

變數名稱	模型係數	T 值	P 值	VIF 值
截距項	-1.25268058	-14.03	<0.01	-
離島服務註記 Offshore_flag	0.13029896	3.65	<0.01	1.00
教育程度年薪中位數 EDU_MID_log	0.20232034	9.24	<0.01	1.85
年齡 AGE	0.00650707	9.93	<0.01	2.14
行業別虛擬變數(軍人) CAREER_CODE_Military	0.27681903	16.36	<0.01	3.30
行業別虛擬變數(警察) CAREER_CODE_Poice	0.18964101	10.15	<0.01	1.94
行業別虛擬變數(教師) CAREER_CODE_Teacher	0.21322123	12.58	<0.01	2.41

變數名稱	模型係數	T 值	P 值	VIF 值
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_Trim	0.00079781	13.74	<0.01	1.31
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_Trim	0.00871279	10.83	<0.01	1.73
信用卡一年期總消費金額 COLUMN_332_NEW_Trim_Trim02	0.00127949	7.56	<0.01	1.39
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.02459910	1.98	0.048	1.11

C60 子模型的 R-squared 值為 0.4767，Adjusted R-squared 值為 0.4744，所有投入變數之模型係數均顯著異於 0。VIF 值均小於 10，變數之間不存在共線性問題。

在業務的應用上，會希望能夠檢視各變數的重要程度，以了解其貢獻性。因此 KPMG 將上述模型進行標準化迴歸，其迴歸係數便能夠揭露其重要程度。結果與變數對應的權重列表如下：

變數名稱	標準化迴歸係數	變數權重
離島服務註記 Offshore_flag	0.0554	2.71%
教育程度年薪中位數 EDU_MID_log	0.1903	9.31%
年齡 AGE	0.2198	10.75%
行業別虛擬變數(軍人) CAREER_CODE_Military	0.4497	21.99%
行業別虛擬變數(警察) CAREER_CODE_Poice	0.2140	10.46%
行業別虛擬變數(教師) CAREER_CODE_Teacher	0.2959	14.47%
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_Trim	0.2378	11.63%
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_Trim	0.2158	10.55%
信用卡一年期總消費金額 COLUMN_332_NEW_Trim_Trim02	0.1351	6.61%
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.0315	1.54%

根據 C60 子模型的模型結果，投入對應的變數即可預測收入距離，後續透過 4.2 節收入距離公式可推導出預測月收入⁶⁹，結果詳見資料集中的「yhat」欄位。

4.11.3 模型效度

下表為子模型分類 C60 的誤差百分比⁷⁰在各區間的樣本佔比以及 MAPE⁷¹統計結果：

誤差百分比	訓練集	測試集	新增測試集 ⁷²	全樣本
-100%~-50%	0.91%	0.70%	0.87%	0.87%
-50%~-20%	10.89%	12.72%	12.48%	11.25%
-20%~-10%	14.29%	15.85%	15.60%	14.60%
-10%~0%	22.65%	21.60%	21.32%	22.44%
0%~10%	19.47%	21.43%	21.32%	19.86%
10%~20%	14.02%	11.32%	11.27%	13.48%
20%~40%	12.89%	11.85%	11.79%	12.68%
40%~50%	1.87%	2.09%	2.08%	1.92%
50%~75%	2.48%	1.92%	1.91%	2.37%
75%~100%	0.26%	0.35%	0.35%	0.28%
>100%	0.26%	0.17%	1.04%	0.24%
樣本整體 MAPE	16.40%	16.21%	17.61%	16.37%
樣本數	2,296	574	577	2,870

下表則是預估月收入減核實月收入誤差之各區間樣本占比：

誤差範圍	訓練集	測試集	新增測試集	全樣本
<-10 萬	0.91%	1.05%	1.04%	0.94%
-10 萬~-5 萬	2.09%	1.57%	1.56%	1.99%
-5 萬~-3 萬	3.09%	3.48%	3.29%	3.17%
-3 萬~-2 萬	4.27%	5.40%	5.37%	4.49%

⁶⁹ 依模型估計得到的預測收入距離(\hat{y}_d)會以 $e^{\hat{y}_d + \log(\text{子模型分類下之總體趨勢薪資}[\text{MONTH_INCOME_ALL_WA}])}$ 轉換為預測月收入

⁷⁰ 誤差百分比 = (預測月收入-核實月收入) / 核實月收入

⁷¹ 將「yhat」欄位的預測月收入，與「Monthly_Income」欄位的核實月收入套入 MAPE 的計算公式(詳見 2.2.4 節之模型效度指標)，即可計算出預測月收入與核實月收入間的 MAPE

⁷² 新增測試集為 2023/12/5 依貴行提出需求新增之測試集，擷取樣本範圍是訓練集以外樣本，並排除(1)非信貸客群 (2)學歷國中以下之非適用客群 (3)策略預估大於年收>200 萬之非適用客群 (4)貴行授信政策下不予承做之核實年薪 24 萬以下客群

誤差範圍	訓練集	測試集	新增測試集	全樣本
-2 萬~-1 萬	11.11%	12.20%	12.13%	11.32%
-1 萬~0 萬	27.26%	27.18%	26.86%	27.25%
0 萬~1 萬	28.88%	28.22%	28.08%	28.75%
1 萬~2 萬	14.72%	13.41%	13.34%	14.46%
2 萬~3 萬	4.88%	5.05%	5.20%	4.91%
3 萬~5 萬	2.31%	2.09%	2.77%	2.26%
5 萬~10 萬	0.48%	0.35%	0.35%	0.45%
>10 萬	0.00%	0.00%	1.04%	0.00%
平均絕對誤差(元)	13,508.79	13,334.16	13,969.48	13,473.86

由上表結果顯示，C60 子模型大約有 96% 的樣本估計誤差介於 $\pm 50\%$ 之間，而平均絕對誤差⁷³約為 13,509 元。每筆樣本的誤差百分比之詳細結果詳見資料集中的「PerErr」欄位，誤差值詳見「Err」欄位。

■ 各收入區間⁷⁴

下表是將全樣本分隔成多個預估年收入區間，並計算在各區間樣本下不同誤差範圍下的樣本占比，以及各區間的樣本數、平均絕對誤差(MAE)、平均絕對誤差百分比(MAPE)。後續的應用策略框架報告書將會再說明 C60 子模型與分量迴歸模型的搭配使用。

誤差範圍 (萬元)	預估年收入區間(萬元)												總計
	0-30	30-40	40-50	50-75	75-100	100-150	150-200	200-250	250-300	300-400	400-500	> 500	
<-10	0.0%	0.0%	0.0%	0.0%	0.2%	0.5%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.9%
-10~-5	0.0%	0.0%	0.0%	0.0%	0.5%	1.2%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	2.0%
-5~-3	0.0%	0.0%	0.0%	0.5%	1.2%	1.2%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	3.2%
-3~-2	0.0%	0.0%	0.0%	1.0%	2.0%	1.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.5%
-2~-1	0.0%	0.0%	0.0%	3.6%	4.2%	3.4%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	11.3%
-1~0	0.0%	0.0%	0.1%	13.4%	8.6%	4.9%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	27.2%
0~1	0.0%	0.0%	0.0%	14.4%	8.8%	5.0%	0.5%	0.0%	0.0%	0.0%	0.0%	0.0%	28.7%
1~2	0.0%	0.0%	0.0%	4.3%	5.2%	4.7%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	14.5%
2~3	0.0%	0.0%	0.0%	0.4%	1.7%	2.1%	0.6%	0.0%	0.0%	0.0%	0.0%	0.0%	4.9%
3~5	0.0%	0.0%	0.0%	0.0%	0.2%	1.1%	0.8%	0.1%	0.0%	0.0%	0.0%	0.0%	2.3%
5~10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	0.5%
>10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

⁷³ 平均絕對誤差(MAE)即為每一筆樣本預測月收入與核實月收入相減後取絕對值再平均

⁷⁴ 各收入區間是以年收入(Yearly_Income)欄位進行判斷，根據年收入高低劃分成 15 個不同所得級距的區間。

樣本數	0	0	3	1,080	936	740	94	12	3	2	0	0	2,870
MAE	0	0	3,773	7,481	12,895	19,098	34,691	55,035	84,554	100,950	0	0	13,474
MAPE	0.0%	0.0%	10.9%	14.0%	16.7%	17.7%	26.6%	33.1%	39.8%	33.7%	0.0%	0.0%	16.4%

4.11.4 模型變數相關說明

■ 離島服務註記

變數名稱為 Offshore_flag，變數來源為資料集所彙整的「公司郵遞區號(POST_CODE_COMPANY)」欄位。該變數相關處理是依據「特徵變數轉換檔」⁷⁵的地理區域變數總表中所整理的區域分類(AREA_DESC)欄位進行對照。

鑒於軍公教警人員於離島服務會有額外的離島加給，相比條件相同的本島軍公教警人員應有較高的薪資收入，故當公司郵遞區號(POST_CODE_COMPANY)屬區域分類(AREA_DESC)中分類為「Offshore」者賦值為 1，其餘則賦值為 0。

■ 教育程度年薪中位數

變數名稱為 EDU_MID_log，變數來源為資料集所彙整的「教育程度(EDUCATION)」欄位。該變數相關處理是以行政院主計總處於 110 年所發布之「工業及服務業受僱員工全年總薪資中位數及分布統計結果」⁷⁶中的表 1-3「工業及服務業受僱員工全年總薪資中位數」統計之年薪中位數(單位：萬元)作為薪資水準進行套用，資料期間共包含了 105~110 年度，KPMG 也將數據彙整於「特徵變數轉換檔」⁷⁷的學歷變數總表頁籤中。

考量到該份外部資料更新時點為隔年 12 月底，例如：2021 年底發布的外部資料為 2020 年度的薪資統計，因此 2022 年 1 月至 2022 年 12 月的樣本可套用最新發布之 2020 年度的薪資統計，故以下列規則衍生變數：

- APPROVE_DATE 在 2018/1/1~2018/12/31，依據 EDUCATION 所對應的學歷套用 105 年度薪資統計，命名為 EDU_MID_2016
- APPROVE_DATE 在 2019/1/1~2019/12/31，依據 EDUCATION 所對應的學歷套用 106 年度薪資統計，命名為 EDU_MID_2017

⁷⁵ 「特徵變數轉換檔」的地理區域變數總表請參考附錄 5.6 之 EXCEL 檔

⁷⁶ 「工業及服務業受僱員工全年總薪資中位數及分布統計結果」請參考附錄 5.5 之 PDF 檔

⁷⁷ 「特徵變數轉換檔」的學歷變數總表請參考附錄 5.6 之 EXCEL 檔

- APPROVE_DATE 在 2020/1/1~2020/12/31，依據 EDUCATION 所對應的學歷套用 107 年度薪資統計，命名為 EDU_MID_2018
- APPROVE_DATE 在 2021/1/1~2021/12/31，依據 EDUCATION 所對應的學歷套用 108 年度薪資統計，命名為 EDU_MID_2019
- APPROVE_DATE 在 2022/1/1~2022/12/31，依據 EDUCATION 所對應的學歷套用 109 年度薪資統計，命名為 EDU_MID_2020
- APPROVE_DATE 在 2023/1/1~2023/12/31，依據 EDUCATION 所對應的學歷套用 110 年度薪資統計，命名為 EDU_MID_2021

與資料集進行 JOIN 後再將 EDU_MID_XXXX 欄位進行整併，並命名為 EDU_MID，若 EDUCATION 缺值則同樣給予總計的中位數薪資水準。最終再將 EDU_MID 取自然對數，並命名為 EDU_MID_log。下圖為部分節錄自附錄 5.6 之特徵變數轉換檔的學歷變數總表：

	A	B	C	D	E	F	G	H
1	EDU_CODE	EDU_CLASS	EDU_MID_2016	EDU_MID_2017	EDU_MID_2018	EDU_MID_2019	EDU_MID_2020	EDU_MID_2021
2		0 研究所	88.4	90.4	94.1	95.2	96	100.2
3		1 大專	51.1	51.8	53.2	53.6	53.7	54.5
4		2 高中	41.2	42.2	43.6	44.2	44.7	44.9
5		3 國中以下	37.3	38	39.4	40.1	40.5	40.7

■ 年齡

變數名稱為 AGE，直接取自資料集的「年齡(AGE)」欄位。

■ 行業別虛擬變數(軍人)

變數名稱為 CAREER_CODE_Military，變數來源為資料集的「CAREER_CODE(行業代碼大類)」欄位，當 CAREER_CODE 等於「O-1」(軍人)時賦值為 1，其他則賦值為 0。

■ 行業別虛擬變數(警察)

變數名稱為 CAREER_CODE_Poice，變數來源為資料集的「CAREER_CODE(行業代碼大類)」欄位，當 CAREER_CODE 等於「O-3」(警察)時賦值為 1，其他則賦值為 0。

■ 行業別虛擬變數(教師)

變數名稱為 CAREER_CODE_Teacher，變數來源為資料集的「CAREER_CODE(行業代碼大類)」欄位，當 CAREER_CODE 等於「P-1」(教師)時賦值為 1，其他則賦值為 0。

■ 無擔保貸款總額度

變數名稱為 VB011_Loan_NonSecur_Limit_NEW_Trim，變數來源為 貴行於 2023/9/19 所提供資料表中的「無擔保貸款總額度(VB011_Loan_NonSecur_Limit)」欄位，後續將該欄位重新命名為 VB011_Loan_NonSecur_Limit_NEW，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 VB011_Loan_NonSecur_Limit_NEW 除以 10,000，並且參考同業相關信貸最高可貸額度普遍設定為 500 萬，故設定 500 萬元作為該變數之上限值，即 $\text{VB011_Loan_NonSecur_Limit_NEW} > 5,000,000$ ，則等於 5,000,000。

■ 近 12 個月平均每發卡銀行永久額度

變數名稱為 VK140_Card_Avg_PerBankLimit_12M_NEW_Trim，變數來源為 貴行於 2023/9/19 所提供資料表中的「近 12 個月平均每發卡銀行永久額度(VK140_Card_Avg_PerBankLimit_12M)」欄位，後續將該欄位重新命名為 VK140_Card_Avg_PerBankLimit_12M_NEW，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 VK140_Card_Avg_PerBankLimit_12M_NEW 除以 10,000，並且參考同業相關信用卡給額設定，發卡額度若超過 50 萬元一般屬於頂級卡客，於額度申請時須提供較多元完整可認列資產在相當程度以上的證明，故設定 50 萬元作為該變數之上限值，即 $\text{VK140_Card_Avg_PerBankLimit_12M_NEW} > 500,000$ ，則等於 500,000。

■ 信用卡一年期總消費金額

變數名稱為 COLUMN_332_NEW_Trim02，變數來源為 貴行於 2023/9/19 所提供資料表中的「信用卡一年期總消費金額(COLUMN_332)」欄位，後續將該欄位重新命名為 COLUMN_332_NEW，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 COLUMN_332_NEW 除以 10,000，並且參考同業最頂級信用卡的 VIP 待遇門檻，常有年消費超過 300 萬元即可加入 VIP 享特殊尊榮禮遇，故設定 300 萬元作為該變數之上限值，即 $\text{COLUMN_332_NEW} > 3,000,000$ ，則等於 3,000,000。

■ 有擔保貸款高額度註記

變數名稱為 VB002_Loan_Secur_Limit_NEW_flag，變數來源為 貴行於 2023/9/19 所提供資料表中的「擔保貸款總額度(VB002_Loan_Secur_Limit)」欄位，後續將該欄位重新命名為 VB002_Loan_Secur_Limit_NEW，空值則補 0。該變數額外相關處理是檢視並參考變數來源欄位的每個百分位點之數值並反覆測試模型找到最佳之高額度門檻，當變數值大於等於門檻值則賦值為 1，小於門檻值則賦值為 0，而最終 C60 子模型之高額度門檻訂為 4,000,000 元。

4.11.5 分量迴歸模型結果

以 C60 子模型投入變數為基礎，進行第 40 百分位數(Q40)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.35322316	-11.71	<0.01
離島服務註記 Offshore_flag	0.14476749	2.61	<0.01
教育程度年薪中位數 EDU_MID_log	0.22500094	7.79	<0.01
年齡 AGE	0.00566270	7.05	<0.01
行業別虛擬變數(軍人) CAREER_CODE_Military	0.29014298	13.65	<0.01
行業別虛擬變數(警察) CAREER_CODE_Poice	0.23956973	9.88	<0.01
行業別虛擬變數(教師) CAREER_CODE_Teacher	0.26865030	11.22	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_Trim	0.00063745	8.72	<0.01
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_Trim	0.00639323	6.41	<0.01
信用卡一年期總消費金額 COLUMN_332_NEW_Trim02	0.00112921	5.22	<0.01
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.02500831	1.94	0.052

以 C60 子模型投入變數為基礎，進行第 35 百分位數(Q35)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.31841663	-12.19	<0.01
離島服務註記 Offshore_flag	0.11144331	1.76	0.078
教育程度年薪中位數 EDU_MID_log	0.19849664	7.35	<0.01
年齡 AGE	0.00652212	9.25	<0.01
行業別虛擬變數(軍人) CAREER_CODE_Military	0.31745175	16.44	<0.01
行業別虛擬變數(警察) CAREER_CODE_Poice	0.24643633	10.84	<0.01
行業別虛擬變數(教師) CAREER_CODE_Teacher	0.27495406	11.53	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_Trim	0.00067396	12.69	<0.01
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_Trim	0.00629829	7.71	<0.01
信用卡一年期總消費金額 COLUMN_332_NEW_Trim_Trim02	0.00107594	5.63	<0.01
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.03093640	2.93	<0.01

以 C60 子模型投入變數為基礎，進行第 30 百分位數(Q30)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.22567287	-10.95	<0.01
離島服務註記 Offshore_flag	0.11658405	5.64	<0.01
教育程度年薪中位數 EDU_MID_log	0.16978199	6.00	<0.01
年齡 AGE	0.00631923	9.33	<0.01
行業別虛擬變數(軍人) CAREER_CODE_Military	0.31638582	16.15	<0.01
行業別虛擬變數(警察) CAREER_CODE_Poice	0.23578389	10.86	<0.01
行業別虛擬變數(教師) CAREER_CODE_Teacher	0.25168032	10.23	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_Trim	0.00075774	16.13	<0.01

變數名稱	模型係數	T 值	P 值
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_Trim	0.00657889	7.83	<0.01
信用卡一年期總消費金額 COLUMN_332_NEW_Trim_Trim02	0.00104275	5.11	<0.01
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.03132471	2.50	0.012

以 C60 子模型投入變數為基礎，進行第 25 百分位數(Q25)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.18293831	-10.47	<0.01
離島服務註記 Offshore_flag	0.12554367	4.79	<0.01
教育程度年薪中位數 EDU_MID_log	0.14842058	5.31	<0.01
年齡 AGE	0.00657070	8.67	<0.01
行業別虛擬變數(軍人) CAREER_CODE_Military	0.33082963	14.48	<0.01
行業別虛擬變數(警察) CAREER_CODE_Poice	0.24324694	10.01	<0.01
行業別虛擬變數(教師) CAREER_CODE_Teacher	0.25000108	9.89	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_Trim	0.00079119	14.77	<0.01
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_Trim	0.00661587	7.87	<0.01
信用卡一年期總消費金額 COLUMN_332_NEW_Trim_Trim02	0.00095769	5.33	<0.01
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.02839976	2.40	0.016

以 C60 子模型投入變數為基礎，進行第 20 百分位數(Q20)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-1.16036786	-11.35	<0.01
離島服務註記 Offshore_flag	0.13899861	7.71	<0.01

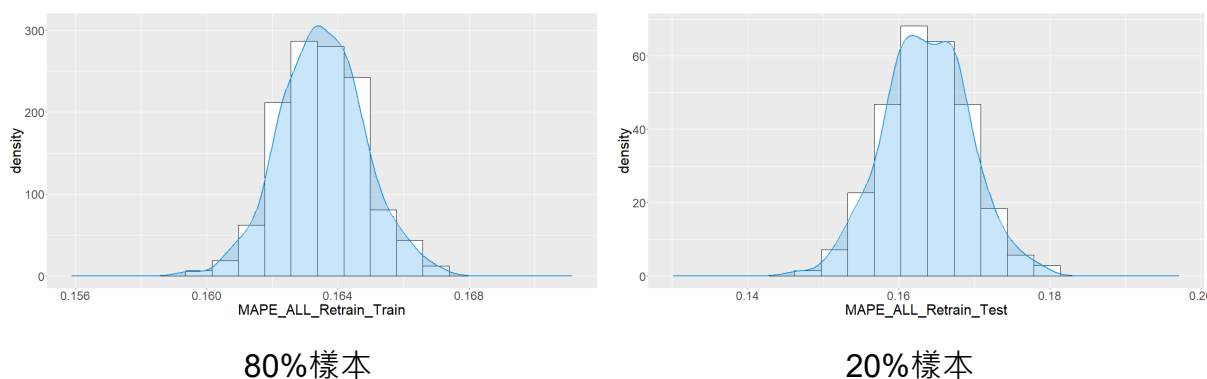
變數名稱	模型係數	T 值	P 值
教育程度年薪中位數 EDU_MID_log	0.12849237	5.09	<0.01
年齡 AGE	0.00645643	8.40	<0.01
行業別虛擬變數(軍人) CAREER_CODE_Military	0.34642062	16.62	<0.01
行業別虛擬變數(警察) CAREER_CODE_Poice	0.25275011	10.23	<0.01
行業別虛擬變數(教師) CAREER_CODE_Teacher	0.25440807	11.16	<0.01
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_Trim	0.00083817	16.89	<0.01
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_Trim	0.00747813	11.07	<0.01
信用卡一年期總消費金額 COLUMN_332_NEW_Trim_Trim02	0.00102545	7.32	<0.01
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.04114928	3.89	<0.01

4.11.6 拔靴法交叉驗證

■ 模型效度穩定性

根據交叉驗證重新擬合結果，訓練集與測試集樣本的 MAPE 分布均集中於平均值，且變異係數小於 30%，顯示 C60 子模型的模型效度穩定性佳。

下圖與下表為 200 次 Bootstrapping 交叉驗證的試驗結果，每次試驗將樣本根據核實所得區間進行分層抽樣，並隨機劃分出 80% 為訓練集，20% 為測試集，重新訓練 C60 子模型後所得到樣本整體 MAPE 的分布如下：



樣本範圍	樣本數	MAPE 平均值	MAPE 標準差	MAPE 變異係數
80%樣本	2,296	16.38%	0.13%	0.78%
20%樣本	574	16.39%	0.56%	3.44%

■ 模型係數穩定性

根據交叉驗證重新擬合結果，各模型係數之變異係數均在 30%以下，顯示 C60 子模型的模型係數估計結果具可靠性。

經 200 次 Bootstrapping 交叉驗證，每次試驗抽出 80%的訓練集樣本並重新訓練 C60 子模型後，所得到的模型係數分布如下表：

變數名稱	平均值	標準差	變異係數
離島服務註記 Offshore_flag	0.1304	1.19%	9.15%
教育程度年薪中位數 EDU_MID_log	0.2193	1.11%	5.05%
年齡 AGE	0.0063	0.03%	5.33%
行業別虛擬變數(軍人) CAREER_CODE_Military	0.2871	0.82%	2.84%
行業別虛擬變數(警察) CAREER_CODE_Poice	0.2024	0.91%	4.47%
行業別虛擬變數(教師) CAREER_CODE_Teacher	0.2188	0.89%	4.09%
無擔保貸款總額度 VB011_Loan_NonSecur_Limit_NEW_Trim	0.0008	0.00%	4.13%
近 12 個月平均每發卡銀行永久額度 VK140_Card_Avg_PerBankLimit_12M_NEW_Trim	0.0086	0.04%	4.90%
信用卡一年期總消費金額 COLUMN_332_NEW_Trim	0.0012	0.01%	8.57%
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.0362	0.60%	16.58%

4.12 非穩收族群子模型多變量分析與模型驗證結果

針對非穩收族群子模型所應用到的資料集說明請參考 4.4 節，而應用到的資料集及結果輸出名稱列示如下：

資料集說明	資料集名稱
訓練集	TrainingData_Unstable_Output.RData
測試集	TestingData_Unstable_Output.RData
建模樣本	ModelData_Unstable_Output.RData

4.12.1 模型設定

$$\begin{aligned} \text{收入距離} = & \alpha + \beta_1 \text{年齡} + \beta_2 \text{戶籍地年薪中位數比率} + \beta_3 \log(\text{教育程度年薪中位數}) \\ & + \beta_4 \log(\text{性別年薪中位數}) + \beta_5 \text{約聘公教人員註記} + \beta_6 \text{小商號負責人註記} \\ & + \beta_7 \text{上市櫃公司溢酬} + \beta_8 \text{無擔保貸款高額度註記} \\ & + \beta_9 \text{近 6 個月平均每發卡銀行永久額度} + \beta_{10} \text{信用卡一年期總消費金額} \\ & + \beta_{11} \text{有擔保貸款高額度註記} + \varepsilon \end{aligned}$$

4.12.2 模型結果

以 6,188 筆訓練集資料配適多元線性迴歸，模型係數四捨五入至小數第 8 位，非穩收族群子模型結果如下：

變數名稱	模型係數	T 值	P 值	VIF 值
截距項	-2.96114677	-10.23	<0.01	-
年齡 AGE	0.00557113	8.13	<0.01	1.25
戶籍地年薪中位數比率 RIGIST_MID	0.07342385	3.14	<0.01	1.07
教育程度年薪中位數 EDU_MID_log	0.12045750	4.43	<0.01	1.12
性別年薪中位數 GEN_MID_log	0.45042572	6.20	<0.01	1.06
約聘公教人員註記 Public_flag	0.11967026	6.19	<0.01	1.13
小商號負責人註記 CHAIRMAN_FLAG_Dum	0.39138953	14.78	<0.01	1.11

變數名稱	模型係數	T 值	P 值	VIF 值
上市櫃公司溢酬 Salary_Company_Prem03	0.12625720	10.02	<0.01	1.23
無擔保貸款高額度註記 VB011_Loan_NonSecur_Limit_NEW_flag	0.38340536	27.65	<0.01	1.24
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_Trim	0.01427803	15.86	<0.01	1.76
信用卡一年期總消費金額 COLUMN_332_NEW_Trim02	0.00275887	19.01	<0.01	1.49
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.14655790	9.96	<0.01	1.13

非穩收族群子模型的 R-squared 值為 0.4496，Adjusted R-squared 值為 0.4486，所有投入變數之模型係數均顯著異於 0。VIF 值均小於 10，變數之間不存在共線性問題。

在業務的應用上，會希望能夠檢視各變數的重要程度，以了解其貢獻性。因此 KPMG 將上述模型進行標準化迴歸，其迴歸係數便能夠揭露其重要程度。結果與變數對應的權重列表如下：

變數名稱	標準化迴歸係數	變數權重
年齡 AGE	0.0860	6.40%
戶籍地年薪中位數比率 RIGIST_MID	0.0306	2.28%
教育程度年薪中位數 EDU_MID_log	0.0443	3.30%
性別年薪中位數 GEN_MID_log	0.0603	4.49%
約聘公教人員註記 Public_flag	0.0622	4.63%
小商號負責人註記 CHAIRMAN_FLAG_Dum	0.1471	10.94%
上市櫃公司溢酬 Salary_Company_Prem03	0.1052	7.83%
無擔保貸款高額度註記 VB011_Loan_NonSecur_Limit_NEW_flag	0.2904	21.61%
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_Trim	0.1988	14.79%
信用卡一年期總消費金額	0.2189	16.29%

變數名稱	標準化迴歸係數	變數權重
COLUMN_332_NEW_Trim02		
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.1002	7.45%

根據非穩收族群子模型的模型結果，投入對應的變數即可預測收入距離，後續透過 4.2 節收入距離公式可推導出預測月收入⁷⁸，結果詳見資料集中的「yhat」欄位。

4.12.3 模型效度

下表為子模型分類非穩收族群的誤差百分比⁷⁹在各區間的樣本佔比以及 MAPE⁸⁰統計結果：

誤差百分比	訓練集	測試集	新增測試集 ⁸¹	全樣本
-100%~-50%	5.94%	6.28%	6.24%	6.01%
-50%~-20%	20.98%	20.78%	16.43%	20.94%
-20%~-10%	9.91%	10.68%	8.75%	10.07%
-10%~0%	9.69%	11.13%	9.12%	9.98%
0%~10%	9.67%	9.39%	7.63%	9.61%
10%~20%	8.73%	8.48%	6.77%	8.68%
20%~40%	14.93%	14.11%	11.68%	14.77%
40%~50%	5.43%	4.08%	3.79%	5.16%
50%~75%	8.21%	8.61%	11.89%	8.29%
75%~100%	3.63%	3.62%	7.84%	3.63%
>100%	2.88%	2.85%	9.87%	2.88%
樣本整體 MAPE	33.20%	32.65%	46.17%	33.09%
樣本數	6,174	1,545	1,875	7,719

⁷⁸ 依模型估計得到的預測收入距離(\hat{y}_d)會以 $e^{\hat{y}_d + \log(\text{子模型分類下之總體趨勢薪資}[\text{MONTH_INCOME_ALL_WA}])}$ 轉換為預測月收入

⁷⁹ 誤差百分比 = (預測月收入-核實月收入) / 核實月收入

⁸⁰ 將「yhat」欄位的預測月收入，與「Monthly_Income」欄位的核實月收入套入 MAPE 的計算公式(詳見 2.2.4 節之模型效度指標)，即可計算出預測月收入與核實月收入間的 MAPE

⁸¹ 新增測試集為 2023/12/5 依貴行提出需求新增之測試集，擷取樣本範圍是訓練集以外樣本，並排除(1)非信貸客群 (2) 近 6M 每發卡銀行永久額度>50 萬之非適用客群 (3)年齡≥60 歲之非適用客群 (4)信用卡 1Y 總消費額>300 萬之非適用客群 (5)策略預估大於年收>200 萬之非適用客群 (6)貴行授信政策下不予承做之核實年薪 24 萬以下客群

下表則是預估月收入減核實月收入誤差之各區間樣本占比：

誤差範圍	訓練集	測試集	新增測試集	全樣本
<-10 萬	3.39%	3.88%	3.89%	3.48%
-10 萬~-5 萬	6.69%	6.41%	5.12%	6.63%
-5 萬~-3 萬	6.64%	6.86%	5.55%	6.68%
-3 萬~-2 萬	6.11%	5.50%	4.37%	5.99%
-2 萬~-1 萬	9.14%	10.42%	8.59%	9.39%
-1 萬~0 萬	14.56%	15.79%	13.01%	14.81%
0 萬~1 萬	20.18%	20.32%	17.07%	20.21%
1 萬~2 萬	17.62%	15.60%	20.69%	17.22%
2 萬~3 萬	7.58%	7.96%	12.80%	7.66%
3 萬~5 萬	5.77%	4.85%	6.61%	5.58%
5 萬~10 萬	1.99%	2.27%	2.29%	2.05%
>10 萬	0.34%	0.13%	3.89%	0.30%
平均絕對誤差(元)	25,361.48	25,360.53	29,888.53	25,361.29

由上表結果顯示，非穩收族群子模型大約有 79% 的樣本估計誤差介於 $\pm 50\%$ 之間，而平均絕對誤差⁸² 約為 25,361 元。每筆樣本的誤差百分比之詳細結果詳見資料集中的「PerErr」欄位，誤差值詳見「Err」欄位。

■ 各收入區間⁸³

下表是將全樣本分隔成多個預估年收入區間，並計算在各區間樣本下不同誤差範圍下的樣本占比，以及各區間的樣本數、平均絕對誤差(MAE)、平均絕對誤差百分比(MAPE)。後續的應用策略框架報告書將會再說明非穩收族群子模型與分量迴歸模型的搭配使用。

誤差範圍 (萬元)	預估年收入區間(萬元)												總計
	0-30	30-40	40-50	50-75	75-100	100-150	150-200	200-250	250-300	300-400	400-500	> 500	
<-10	0.0%	0.0%	0.0%	0.6%	0.9%	1.1%	0.6%	0.1%	0.1%	0.0%	0.0%	0.0%	3.5%
-10~-5	0.0%	0.0%	0.2%	1.7%	2.0%	2.1%	0.4%	0.1%	0.0%	0.1%	0.0%	0.0%	6.6%
-5~-3	0.0%	0.0%	0.3%	2.7%	1.6%	1.5%	0.5%	0.1%	0.0%	0.0%	0.0%	0.0%	6.7%
-3~-2	0.0%	0.1%	0.5%	2.7%	1.4%	1.1%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	6.0%
-2~-1	0.0%	0.3%	1.5%	4.6%	1.5%	1.3%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	9.4%
-1~0	0.0%	0.4%	3.4%	7.4%	1.9%	1.4%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	14.8%

⁸² 平均絕對誤差(MAE)即為每一筆樣本預測月收入與核實月收入相減後取絕對值再平均

⁸³ 各收入區間是以年收入(Yearly_Income)欄位進行判斷，根據年收入高低劃分成 15 個不同所得級距的區間。

0~1	0.0%	0.3%	4.5%	10.8%	2.6%	1.6%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	20.2%
1~2	0.0%	0.0%	1.7%	10.9%	2.5%	1.7%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	17.2%
2~3	0.0%	0.0%	0.0%	3.4%	2.3%	1.7%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	7.7%
3~5	0.0%	0.0%	0.0%	0.3%	2.0%	2.5%	0.6%	0.1%	0.0%	0.0%	0.0%	0.0%	5.6%
5~10	0.0%	0.0%	0.0%	0.0%	0.1%	0.9%	0.6%	0.3%	0.1%	0.0%	0.0%	0.0%	2.0%
>10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.1%	0.0%	0.0%	0.3%
樣本數	0	83	927	3,478	1,458	1,298	325	88	31	25	6	0	7,719
MAE	0	8,377	9,849	17,185	32,360	39,233	55,134	61,030	77,242	73,684	90,000	0	25,361
MAPE	0.0%	20.8%	22.7%	30.9%	39.2%	38.1%	36.1%	51.4%	41.6%	31.2%	46.8%	0.0%	33.1%

4.12.4 模型變數相關說明

■ 年齡

變數名稱為 AGE，直接取自資料集的「年齡(AGE)」欄位。

■ 戶籍地年薪中位數比率

變數名稱為 RIGIST_MID，變數來源為資料集所彙整的「戶籍郵遞區號(POST_CODE_REGISTER)」與「現居郵遞區號(POST_CODE_CONTACT)」欄位。該變數相關處理是根據財政部統計處公布之「我國各縣市及鄉鎮市區綜合所得稅結算申報情形」⁸⁴，選取各鄉鎮市區的「綜合所得總額-中位數」(單位：萬元)作為薪資水準，但統計資料說明中表示：自 107 年起不含分離課稅股利所得、108 年起不含可扣除薪資必要費用，因此每年的薪資水準無法直接以絕對金額進行跨期比較，故將各鄉鎮市區所得皆除以所得最低的鄉鎮市區轉換成比率後進行套用，資料期間共包含了 104~109 年度，KPMG 也將其數據彙整於「特徵變數轉換檔」⁸⁵的地理區域變數總表頁籤中。

考量到該年的個人所得於次年課稅，且該份外部資料更新時點約為課徵年度的隔年 8 月公布，故以下列規則衍生變數：

- APPROVE_DATE 在 2017/9/1~2018/8/31，依據 POST_CODE_REGISTER 所在縣市套用 104 年度薪資統計，命名為 RIGIST_MID_2015
- APPROVE_DATE 在 2018/9/1~2019/8/31，依據 POST_CODE_REGISTER 所在縣市套用 105 年度薪資統計，命名為 RIGIST_MID_2016

⁸⁴ 財政部統計處公布之「我國各縣市及鄉鎮市區綜合所得稅結算申報情形」資料請參考以下網址：

<https://service.mof.gov.tw/public/Data/statistic/d3.js/demo/tax02/index.html>

⁸⁵ 「特徵變數轉換檔」的地理區域變數總表請參考附錄 5.6 之 EXCEL 檔

- APPROVE_DATE 在 2019/9/1~2020/8/31，依據 POST_CODE_REGISTER 所在縣市套用 106 年度薪資統計，命名為 RIGIST_MID_2017
- APPROVE_DATE 在 2020/9/1~2021/8/31，依據 POST_CODE_REGISTER 所在縣市套用 107 年度薪資統計，命名為 RIGIST_MID_2018
- APPROVE_DATE 在 2021/9/1~2022/8/31，依據 POST_CODE_REGISTER 所在縣市套用 108 年度薪資統計，命名為 RIGIST_MID_2019
- APPROVE_DATE 在 2022/9/1~2023/8/31，依據 POST_CODE_REGISTER 所在縣市套用 109 年度薪資統計，命名為 RIGIST_MID_2020

惟新竹市與嘉義市之郵遞區號並未細分至區，故改採縣市合計額，另有部分外島地區未有相關薪資統計，故採最低的薪資水準。與資料集進行 JOIN 後再將 RIGIST_MID_XXXX 欄位進行整併，並命名為 RIGIST_MID，若戶籍郵遞區號(POST_CODE_REGISTER)缺值，則優先以現居郵遞區號(POST_CODE_CONTACT)賦值，若兩者都缺則將戶籍地年薪中位數比率設定為 1。下圖為部分節錄自附錄 5.6 之特徵變數轉換檔的地理區域變數總表：

	A	B	H	I	J	K	L
1	ZIP_CODE	ZIP_DESC	RIGIST_MID_2015	RIGIST_MID_2016	RIGIST_MID_2017	RIGIST_MID_2018	RIGIST_MID_2019
2	100	臺北市中正區	1.753191489	1.702083333	1.704166667	1.6861167	2.167763158
3	103	臺北市大同區	1.378723404	1.358333333	1.366666667	1.368209256	1.592105263
4	104	臺北市中山區	1.555319149	1.522916667	1.533333333	1.515090543	1.865131579
5	105	臺北市松山區	1.738297872	1.695833333	1.691666667	1.672032193	2.101973684
6	106	臺北市大安區	1.844680851	1.795833333	1.785416667	1.770623742	2.315789474
7	108	臺北市萬華區	1.30212766	1.28125	1.283333333	1.283702213	1.467105263
8	110	臺北市信義區	1.557446809	1.51875	1.522916667	1.507042254	1.809210526
9	111	臺北市士林區	1.495744681	1.45625	1.45625	1.45472837	1.733552632
10	112	臺北市北投區	1.485106383	1.45625	1.466666667	1.458752515	1.730263158
11	114	臺北市內湖區	1.653191489	1.610416667	1.616666667	1.61971831	1.950657895
12	115	臺北市南港區	1.457446809	1.4375	1.44375	1.456740443	1.700657895
13	116	臺北市文山區	1.638297872	1.591666667	1.591666667	1.579476861	1.904605263
14	200	基隆市仁愛區	1.242553191	1.2	1.2	1.207243461	1.365131579
15	201	基隆市信義區	1.285106383	1.254166667	1.24375	1.251509054	1.368421053
16	202	基隆市中正區	1.208510638	1.189583333	1.179166667	1.175050302	1.276315789
17	203	基隆市中山區	1.217021277	1.202083333	1.19375	1.189134809	1.282894737
18	204	基隆市安樂區	1.329787234	1.289583333	1.289583333	1.29778672	1.427631579
19	205	基隆市暖暖區	1.287234043	1.247916667	1.2375	1.207243461	1.302631579
20	206	基隆市七堵區	1.219148936	1.189583333	1.19375	1.179074447	1.266447368
21	207	新北市萬里區	1.078723404	1.058333333	1.077083333	1.084507042	1.108552632
22	208	新北市金山區	1.061702128	1.03125	1.05	1.038229376	1.065789474
23	220	新北市板橋區	1.329787234	1.310416667	1.316666667	1.327967807	1.519736842
24	221	新北市汐止區	1.368085106	1.325	1.322916667	1.329979879	1.486842105
25	222	新北市深坑區	1.29787234	1.260416667	1.26875	1.267605634	1.348684211
26	223	新北市石碇區	1.134042553	1.079166667	1.104166667	1.09054326	1.108552632
27	224	新北市瑞芳區	1.09787234	1.095833333	1.089583333	1.08249497	1.111842105
28	226	新北市平溪區	1.087234043	1.091666667	1.097916667	1.088531187	1.098684211
29	227	新北市雙溪區	1.125531915	1.110416667	1.104166667	1.072434608	1.131578947
30	228	新北市貢寮區	1.04893617	1.05	1.070833333	1.054325956	1.072368421

■ 教育程度年薪中位數

變數名稱為 EDU_MID_log，變數來源為資料集所彙整的「教育程度(EDUCATION)」欄位。該變數相關處理是以行政院主計總處於 110 年所發布之「工業及服務業受僱員工全年總薪資中位數及分布統計結果」⁸⁶中的表 1-3「工業及服務業受僱員工全年總薪資中位數」統計之年薪中位數(單位：萬元)作為薪資水準進行套用，資料期間共包含了 105~110 年度，KPMG 也將數據彙整於「特徵變數轉換檔」⁸⁷的學歷變數總表頁籤中。

考量到該份外部資料更新時點為隔年 12 月底，例如：2021 年底發布的外部資料為 2020 年度的薪資統計，因此 2022 年 1 月至 2022 年 12 月的樣本可套用最新發布之 2020 年度的薪資統計，故以下列規則衍生變數：

- APPROVE_DATE 在 2018/1/1~2018/12/31，依據 EDUCATION 所對應的學歷套用 105 年度薪資統計，命名為 EDU_MID_2016
- APPROVE_DATE 在 2019/1/1~2019/12/31，依據 EDUCATION 所對應的學歷套用 106 年度薪資統計，命名為 EDU_MID_2017
- APPROVE_DATE 在 2020/1/1~2020/12/31，依據 EDUCATION 所對應的學歷套用 107 年度薪資統計，命名為 EDU_MID_2018
- APPROVE_DATE 在 2021/1/1~2021/12/31，依據 EDUCATION 所對應的學歷套用 108 年度薪資統計，命名為 EDU_MID_2019
- APPROVE_DATE 在 2022/1/1~2022/12/31，依據 EDUCATION 所對應的學歷套用 109 年度薪資統計，命名為 EDU_MID_2020
- APPROVE_DATE 在 2023/1/1~2023/12/31，依據 EDUCATION 所對應的學歷套用 110 年度薪資統計，命名為 EDU_MID_2021

與資料集進行 JOIN 後再將 EDU_MID_XXXX 欄位進行整併，並命名為 EDU_MID，若 EDUCATION 缺值則同樣給予總計的中位數薪資水準。最終再將 EDU_MID 取自然對數，並命名為 EDU_MID_log。下圖為部分節錄自附錄 5.6 之特徵變數轉換檔的學歷變數總表：

	A	B	C	D	E	F	G	H
1	EDU_CODE	EDU_CLASS	EDU_MID_2016	EDU_MID_2017	EDU_MID_2018	EDU_MID_2019	EDU_MID_2020	EDU_MID_2021
2		0 研究所	88.4	90.4	94.1	95.2	96	100.2
3		1 大專	51.1	51.8	53.2	53.6	53.7	54.5
4		2 高中	41.2	42.2	43.6	44.2	44.7	44.9
5		3 國中以下	37.3	38	39.4	40.1	40.5	40.7

⁸⁶ 「工業及服務業受僱員工全年總薪資中位數及分布統計結果」請參考附錄 5.5 之 PDF 檔

⁸⁷ 「特徵變數轉換檔」的學歷變數總表請參考附錄 5.6 之 EXCEL 檔

■ 性別年薪中位數

變數名稱為 GEN_MID_log，變數來源為資料集所彙整的「性別(gender)」欄位。該變數相關處理是以行政院主計總處於 110 年所發布之「工業及服務業受僱員工全年總薪資中位數及分布統計結果」⁸⁸中的表 1-3「工業及服務業受僱員工全年總薪資中位數」統計之年薪中位數(單位：萬元)作為薪資水準進行套用，資料期間共包含了 105~110 年度，KPMG 也將數據彙整於「特徵變數轉換檔」⁸⁹的性別變數總表頁籤中。

考量到該份外部資料更新時點為隔年 12 月底，例如：2021 年底發布的外部資料為 2020 年度的薪資統計，因此 2022 年 1 月至 2022 年 12 月的樣本可套用最新發布之 2020 年度的薪資統計，故以下列規則衍生變數：

- APPROVE_DATE 在 2018/1/1~2018/12/31，依據 gender 所對應的性別套用 105 年度薪資統計，命名為 GEN_MID_2016
- APPROVE_DATE 在 2019/1/1~2019/12/31，依據 gender 所對應的性別套用 106 年度薪資統計，命名為 GEN_MID_2017
- APPROVE_DATE 在 2020/1/1~2020/12/31，依據 gender 所對應的性別套用 107 年度薪資統計，命名為 GEN_MID_2018
- APPROVE_DATE 在 2021/1/1~2021/12/31，依據 gender 所對應的性別套用 108 年度薪資統計，命名為 GEN_MID_2019
- APPROVE_DATE 在 2022/1/1~2022/12/31，依據 gender 所對應的性別套用 109 年度薪資統計，命名為 GEN_MID_2020
- APPROVE_DATE 在 2023/1/1~2023/12/31，依據 gender 所對應的性別套用 110 年度薪資統計，命名為 GEN_MID_2021

與資料集進行 JOIN 後再將 GEN_MID_XXXX 欄位進行整併，並命名為 GEN_MID，若 gender 缺值則同樣給予總計的中位數薪資水準。最終再將 GEN_MID 取自然對數，並命名為 GEN_MID_log。下圖為部分節錄自附錄 5.6 之特徵變數轉換檔的性別變數總表：

	A	B	C	D	E	F	G	H
1	GEN_CODE	GEN_CLASS	GEN_MID_2016	GEN_MID_2017	GEN_MID_2018	GEN_MID_2019	GEN_MID_2020	GEN_MID_2021
2		1 男	49.7	50.7	52.1	53	53.2	54.9
3		2 女	43.1	44.1	45.6	46.2	46.8	47.3

⁸⁸ 「工業及服務業受僱員工全年總薪資中位數及分布統計結果」請參考附錄 5.5 之 PDF 檔

⁸⁹ 「特徵變數轉換檔」的學歷變數總表請參考附錄 5.6 之 EXCEL 檔

■ 約聘公教人員註記

變數名稱為 `Public_flag`，變數來源為資料集的「`CAREER_CODE`(行業代碼大類)」欄位，當 `CAREER_CODE` 為"O-1"(軍人)、"O-2"(公務人員)、"O-3"(警察)、"P-1"(教師)任一者時賦值為 1，其他則賦值為 0。

■ 小商號負責人註記

變數名稱為 `CHAIRMAN_FLAG_Dum`，變數來源為資料集的「`CHAIRMAN_FLAG`(負責人註記)」欄位，由於非穩收族群子模型中出現的負責人均已定義為小商號負責人，故當 `CHAIRMAN_FLAG` 不為空則表示為小商號負責人，賦值為 1，其他則賦值為 0。

■ 上市櫃公司溢酬

變數名稱為 `Salary_Company_Prem03`，變數來源為資料集所彙整的「公司名稱(`COMPANY`)」、「主計處薪資水準(`MONTH_INCOME_CAREERCODE`)」、「子模型分類總體趨勢薪資(`MONTH_INCOME_ALL_WA`)」欄位。

首先，須將資料處理篇所整理的「上市櫃公司名單列表」與資料集進行 JOIN，若資料集公司名稱(`COMPANY`)欄位中含有「上市櫃公司名單列表」所列公司名稱關鍵字串者，則賦予該樣本對應的公司代碼(`Symbol`)，惟派遣、工讀生等職位之敘薪與其他正職員工相異甚大，其薪資水準不宜視同上市櫃公司之正職員工，故當 `OC_CODE_REVIEW` 為 28, 29, 30, 31, 32, 33, 34, 35, 37, 55, 63, 67, 68 時，公司代碼(`Symbol`)欄位為空。

其次，再依據公司代碼(`Symbol`)以及資料對照年月(`Mapping_YM`)，與「上市櫃員工薪資對照表」進行匹配，即當公司代碼(`Symbol`)相同且資料對照年月(`Mapping_YM`)介於套用起點與套用終點時，給予「非擔任主管職務之全時員工資訊_員工薪資_平均數_仟元每人(`Salary_Company`)」欄位的值作為其所屬公司的薪資水準，並將其乘以 1,000 再除以 12 轉換為平均月收入(單位：元)。

最後，設定一個新的 `Salary_Company_Prem` 欄位，當該客戶為上市櫃公司正職員工，即 `Salary_Company` 非空值時，則賦值為 `Salary_Company` 當作該公司薪資水準；反之，當該客戶非屬上市櫃公司正職員工時，則賦值為該行業的主計處薪資水準(`MONTH_INCOME_CAREERCODE`)作為其所屬產業的薪資水準。最終再將 `Salary_Company_Prem` 除以子模型分類總體趨勢薪資(`MONTH_INCOME_ALL_WA`)，即可得到上市櫃公司溢酬。

■ 無擔保貸款高額度註記

變數名稱為 VB011_Loan_NonSecur_Limit_NEW_flag，變數來源為 貴行於 2023/9/19 所提供資料表中的「無擔保貸款總額度(VB011_Loan_NonSecur_Limit)」欄位，後續將該欄位重新命名為 VB011_Loan_NonSecur_Limit_NEW，空值則補 0。該變數額外相關處理是檢視並參考變數來源欄位的每個百分位點之數值並反覆測試模型找到最佳之高額度門檻，另考量約聘僱人員與其他傭金收入人員在本模型中對此變數的鑑別力較高，故設定當變數值大於等於門檻值且 CHAIRMAN_FLAG_Dum 為 0 則賦值為 1，其餘則賦值為 0，而最終非穩收族群子模型之高額度門檻訂為 1,000,000 元。

■ 近 6 個月平均每發卡銀行永久額度

變數名稱為 VK138_Card_Avg_PerBankLimit_6M_NEW_Trim，變數來源為 貴行於 2023/9/19 所提供資料表中的「近 6 個月平均每發卡銀行永久額度(VK138_Card_Avg_PerBankLimit_6M)」欄位，後續將該欄位重新命名為 VK138_Card_Avg_PerBankLimit_6M_NEW，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 VK138_Card_Avg_PerBankLimit_6M_NEW 除以 10,000，並且參考同業相關信用卡給額設定，發卡額度若超過 50 萬元一般屬於頂級卡客，於額度申請時須提供較多元完整可認列資產在相當程度以上的證明，故設定 50 萬元作為該變數之上限值，即 VK138_Card_Avg_PerBankLimit_6M_NEW > 500,000，則等於 500,000。

■ 信用卡一年期總消費金額

變數名稱為 COLUMN_332_NEW_Trim02，變數來源為 貴行於 2023/9/19 所提供資料表中的「信用卡一年期總消費金額(COLUMN_332)」欄位，後續將該欄位重新命名為 COLUMN_332_NEW，空值則補 0。該變數額外相關處理是將欄位單位進行調整從原本的元轉換為萬元，即 COLUMN_332_NEW 除以 10,000，並且參考同業最頂級信用卡的 VIP 待遇門檻，常有年消費超過 300 萬元即可加入 VIP 享特殊尊榮禮遇，故設定 300 萬元作為該變數之上限值，即 COLUMN_332_NEW > 3,000,000，則等於 3,000,000。

■ 有擔保貸款高額度註記

變數名稱為 VB002_Loan_Secur_Limit_NEW_flag，變數來源為 貴行於 2023/9/19 所提供資料表中的「擔保貸款總額度(VB002_Loan_Secur_Limit)」欄位，後續將該欄位重

新命名為 VB002_Loan_Secur_Limit_NEW，空值則補 0。該變數額外相關處理是檢視並參考變數來源欄位的每個百分位點之數值並反覆測試模型找到最佳之高額度門檻，當變數值大於等於門檻值則賦值為 1，小於門檻值則賦值為 0，而最終非穩收族群子模型之高額度門檻訂為 3,000,000 元。

4.12.5 分量迴歸模型結果

以非穩收族群子模型投入變數為基礎，進行第 40 百分位數(Q40)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-3.44663002	-10.55	<0.01
年齡 AGE	0.00564919	6.55	<0.01
戶籍地年薪中位數比率 RIGIST_MID	0.08432213	2.76	<0.01
教育程度年薪中位數 EDU_MID_log	0.12355863	3.95	<0.01
性別年薪中位數 GEN_MID_log	0.52785007	6.46	<0.01
約聘公教人員註記 Public_flag	0.17750215	10.49	<0.01
小商號負責人註記 CHAIRMAN_FLAG_Dum	0.34734597	5.67	<0.01
上市櫃公司溢酬 Salary_Company_Prem03	0.11296560	7.69	<0.01
無擔保貸款高額度註記 VB011_Loan_NonSecur_Limit_NEW_flag	0.37990181	23.07	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_Trim	0.01709075	16.37	<0.01
信用卡一年期總消費金額 COLUMN_332_NEW_Trim02	0.00281336	15.25	<0.01
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.12198636	7.06	<0.01

以非穩收族群子模型投入變數為基礎，進行第 35 百分位數(Q35)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-3.21611707	-10.92	<0.01

變數名稱	模型係數	T 值	P 值
年齡 AGE	0.00434153	5.32	<0.01
戶籍地年薪中位數比率 RIGIST_MID	0.04291233	1.49	0.14
教育程度年薪中位數 EDU_MID_log	0.13719414	4.84	<0.01
性別年薪中位數 GEN_MID_log	0.46473094	6.32	<0.01
約聘公教人員註記 Public_flag	0.19821994	13.39	<0.01
小商號負責人註記 CHAIRMAN_FLAG_Dum	0.30371447	4.96	<0.01
上市櫃公司溢酬 Salary_Company_Prem03	0.12103007	9.00	<0.01
無擔保貸款高額度註記 VB011_Loan_NonSecur_Limit_NEW_flag	0.37469773	23.80	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_Trim	0.01776847	16.31	<0.01
信用卡一年期總消費金額 COLUMN_332_NEW_Trim02	0.00270196	14.46	<0.01
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.13501260	9.19	<0.01

以非穩收族群子模型投入變數為基礎，進行第 30 百分位數(Q30)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-3.05268867	-10.95	<0.01
年齡 AGE	0.00378122	5.23	<0.01
戶籍地年薪中位數比率 RIGIST_MID	0.05873708	2.57	<0.01
教育程度年薪中位數 EDU_MID_log	0.13031001	4.68	<0.01
性別年薪中位數 GEN_MID_log	0.41787344	6.05	<0.01
約聘公教人員註記 Public_flag	0.21376819	16.30	<0.01
小商號負責人註記 CHAIRMAN_FLAG_Dum	0.25730054	4.55	<0.01
上市櫃公司溢酬 Salary_Company_Prem03	0.12851912	10.02	<0.01

變數名稱	模型係數	T 值	P 值
無擔保貸款高額度註記 VB011_Loan_NonSecur_Limit_NEW_flag	0.38135800	26.30	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_Trim	0.01616680	14.30	<0.01
信用卡一年期總消費金額 COLUMN_332_NEW_Trim02	0.00275001	15.85	<0.01
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.14499964	10.39	<0.01

以非穩收族群子模型投入變數為基礎，進行第 25 百分位數(Q25)的分量迴歸配適，模型結果如下：

變數名稱	模型係數	T 值	P 值
截距項	-2.96375123	-9.61	<0.01
年齡 AGE	0.00304467	3.86	<0.01
戶籍地年薪中位數比率 RIGIST_MID	0.05204105	1.93	0.05
教育程度年薪中位數 EDU_MID_log	0.11596681	3.86	<0.01
性別年薪中位數 GEN_MID_log	0.41058959	5.37	<0.01
約聘公教人員註記 Public_flag	0.23572080	16.12	<0.01
小商號負責人註記 CHAIRMAN_FLAG_Dum	0.20073086	3.08	<0.01
上市櫃公司溢酬 Salary_Company_Prem03	0.12272370	8.67	<0.01
無擔保貸款高額度註記 VB011_Loan_NonSecur_Limit_NEW_flag	0.37741122	21.91	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_Trim	0.01542438	11.91	<0.01
信用卡一年期總消費金額 COLUMN_332_NEW_Trim02	0.00275094	13.75	<0.01
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.15364666	8.28	<0.01

以非穩收族群子模型投入變數為基礎，進行第 20 百分位數(Q20)的分量迴歸配適，模型結果如下：

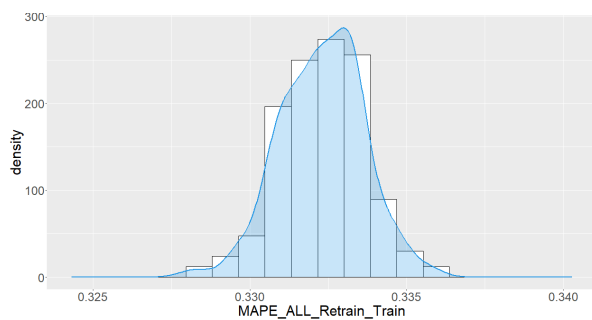
變數名稱	模型係數	T 值	P 值
截距項	-2.83986036	-9.13	<0.01
年齡 AGE	0.00262719	3.18	<0.01
戶籍地年薪中位數比率 RIGIST_MID	0.05765989	2.12	0.03
教育程度年薪中位數 EDU_MID_log	0.12233542	3.97	<0.01
性別年薪中位數 GEN_MID_log	0.36577048	4.74	<0.01
約聘公教人員註記 Public_flag	0.25256802	17.26	<0.01
小商號負責人註記 CHAIRMAN_FLAG_Dum	0.12962390	2.64	<0.01
上市櫃公司溢酬 Salary_Company_Prem03	0.12522358	8.30	<0.01
無擔保貸款高額度註記 VB011_Loan_NonSecur_Limit_NEW_flag	0.37348405	21.97	<0.01
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_Trim	0.01301834	10.04	<0.01
信用卡一年期總消費金額 COLUMN_332_NEW_Trim02	0.00277616	12.09	<0.01
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.14256044	7.71	<0.01

4.12.6 拔靴法交叉驗證

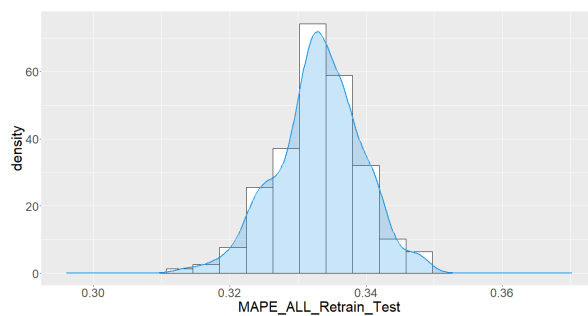
■ 模型效度穩定性

根據交叉驗證重新擬合結果，訓練集與測試集樣本的 MAPE 分布均集中於平均值，且變異係數小於 30%，顯示非穩收族群子模型的模型效度穩定性佳。

下圖與下表為 200 次 Bootstrapping 交叉驗證的試驗結果，每次試驗將樣本根據核實所得區間進行分層抽樣，並隨機劃分出 80% 為訓練集，20% 為測試集，重新訓練非穩收族群子模型後所得到樣本整體 MAPE 的分布如下：



80%樣本



20%樣本

樣本範圍	樣本數	MAPE 平均值	MAPE 標準差	MAPE 變異係數
80%樣本	6,174	33.18%	0.13%	0.40%
20%樣本	1,545	33.28%	0.62%	1.87%

■ 模型係數穩定性

根據交叉驗證重新擬合結果，各模型係數之變異係數均在 30%以下，顯示非穩收族群子模型的模型係數估計結果具可靠性。

經 200 次 Bootstrapping 交叉驗證，每次試驗抽出 80%的訓練集樣本並重新訓練非穩收族群子模型後，所得到的模型係數分布如下表：

變數名稱	平均值	標準差	變異係數
年齡 AGE	0.0058	0.03%	5.75%
戶籍地年薪中位數比率 RIGIST_MID	0.0900	1.11%	12.30%
教育程度年薪中位數 EDU_MID_log	0.1149	1.18%	10.26%
性別年薪中位數 GEN_MID_log	0.4816	2.91%	6.05%
約聘公教人員註記 Public_flag	0.1199	0.63%	5.25%
小商號負責人註記 CHAIRMAN_FLAG_Dum	0.3856	1.72%	4.47%
上市櫃公司溢酬 Salary_Company_Prem03	0.1324	0.60%	4.57%
無擔保貸款高額度註記 VB011_Loan_NonSecur_Limit_NEW_flag	0.3898	0.60%	1.54%
近 6 個月平均每發卡銀行永久額度 VK138_Card_Avg_PerBankLimit_6M_NEW_Trim	0.0138	0.05%	3.55%
信用卡一年期總消費金額	0.0029	0.01%	2.76%

變數名稱	平均值	標準差	變異係數
COLUMN_332_NEW_Trim			
有擔保貸款高額度註記 VB002_Loan_Secur_Limit_NEW_flag	0.1296	0.71%	5.45%

5. 附錄

5.1 行業與職稱別對照表



行業與職稱別對照
表.xlsx

5.2 修訂行業別對照表



修訂行業別對照表.
xlsx

5.3 勞動部職業名稱及工作說明



勞動部職業名稱及
工作說明.pdf

5.4 109 年各縣市工業及服務業企業僱用員工薪資統計 說明



109各縣市工業及
服務業企業僱用員工

5.5 110 年工業及服務業受僱員工全年總薪資中位數及 分布統計結果



110年工業及服務
業受僱員工全年總薪資

5.6 特徵變數轉換檔



5.6

特徵變數轉換.xlsx

5.7 模型效度一覽



5.7

模型效度一覽.xlsx

獨立性規範聲明

本文所提及之一部分或全部服務，依相關獨立性規範，可能無法對 KPMG 之審計客戶及其關係企業提供服務。

聯絡我們

陳世雄 執行副總

財務風險管理服務

T +886 (2) 8101-6666 分機 10940

E schen26@kpmg.com.tw

李鑑剛 副總經理

財務風險管理服務

T +886 (2) 8101-6666 分機 14638

E gklee@kpmg.com.tw

何冠儒 經理

財務風險管理服務

T +886 (2) 8101-6666 分機 15829

E joshuah@kpmg.com.tw

林育正 副理

財務風險管理服務

T +886 (2) 8101-6666 分機 18369

E ivanlin@kpmg.com.tw

林展源 高級顧問師

財務風險管理服務

T +886 (2) 8101-6666 分機 19076

E ryanlin1@kpmg.com.tw

kpmg.com/tw

© 2023 KPMG Advisory Services Co., Ltd., a Taiwan company limited by shares and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

The KPMG name and logo are trademarks used under license by the independent member firms of the KPMG global organization.