



AI 於財稅之應用—以租屋市場查核為例

本文以財政部現有及外蒐資料為基礎，擴充可能影響個人間租賃行為之特徵變數，並建立監督式學習之 AI 模型，現階段模型預測準確性可達 7 成。希冀未來可提供稽徵實務查核輔助之用，逐步改善租屋市場不透明，以維護租稅之公平性。

楊子江（財政部統計處研究員）

壹、前言

隨電腦運算效能提升，人工智慧（Artificial Intelligence, AI）已廣泛見於日常生活中，不論是人臉辨識、搜尋結果推薦、自動駕駛，乃至於藝術創作、撰寫報告等生成式 AI。有鑑於此，財政部於 109 年底訂定「運用人工智慧技術提升稅務行政效能作業計畫」，並於隔年建置資料科學工作坊，由賦稅署、財政資訊中心、五區國稅局及財政部統計處（以

下簡稱本處）等單位參與，運用人工智慧雲端平臺（AI PaaS），處理稅務查核實際問題。迄今已完成數個實作主題，如營業人交易關係網絡分析、偵測短漏開統一發票異常案件等。本文以本處目前執行中主題，就現階段之初步成果加以說明。

貳、實作主題

個人間房屋租賃交易資訊因多數毋須強制揭露¹，往往造成房客處相對弱勢，縱然政府

提供租屋補貼、租金支出扣除額等機制，多仍未加以利用，導致其權利遭剝削，亦造成國稅局不易勾稽，增加查核困難度，且房租短漏報除造成稅基流失，亦不符量能課稅原則，影響租稅公平性。

根據綜合所得稅結算申報統計專冊顯示，108 年僅 2.6 萬戶申報租金支出列舉扣除額，但實價登錄與民間租屋平臺顯示同期約有近 80 萬戶租屋需求²，且同年大專院校學生校外租屋人數即達 26.7 萬人，

顯示租屋市場之需求與資訊存有顯著差異。本實作主題目標為透過 AI 技術之應用，建立漏報租賃所得偵測模型，希冀提高現行漏報租賃所得收入案件識別命中率。

參、資料應用

一、資料運用範圍

本主題以房屋稅稅籍資料為基礎，串接國稅、地方稅及外蒐等相關資料（附表），藉以延展出各方面之多重特徵維度。各表資料於應用上有時間差異，例如房屋稅稅籍資料為當期（t 期），但所得稅須待申報核定後方可取得，故其能應用資料期為前一期（t-1 期）。此外，由於房屋稅稅籍資料包

括多種房屋類型及持有人，為符合本主題目的且考量實際租屋市場需求及模型運用便利性，故僅納入屬個人持有之房屋，且限定由 1 人完全持有，並排除地下室及特殊構造等建物類型。

二、資料處理

為利拓展建模所需特徵維度，另須就資料進行相關處理：

（一）各類財產市價化

考量現行蒐集財產檔案其價值遠低於市價，故另透過外蒐資料加以估計，還原為市價。如房屋財產係利用內政部實價登錄屬於建物交易買賣資料，計算出各縣市每坪市場交易單價中位數，並與房屋財產檔的房屋評定現值中位數比較，換算出各縣市建物價值之市價還原倍數，加以估算。土地財產檔利用內政部公告地價占一般正常交易價格百分比，依照各縣市比重，加以估算土地財產之市場價值。車輛財產估算處理則相對複雜，由於

附表 取用之資料

資料來源	資料名稱	資料期間 (年)	資料用途
財政部	房屋稅中文、課稅主檔、持份人檔	101~108	模型基礎資料
	營業人稅籍月檔	105~110	
	所得大檔（所得彙整檔）	100~107	模型特徵維度
	綜所稅核定檔、核定所得檔	100~107	
	戶政資訊檔	101~108	
	房屋財產檔	101~108	
	土地財產檔	101~108	
	車輛財產檔	101~108	
投資財產檔	101~108		
外部 蒐集	內政部實價登錄資料	101~108	財產價值設算
	591 租屋網站資料	107~108	
	衛福部醫療院所資料	100~108	
	教育部各級學校資料	103~108	
	內政部實價登錄建物買賣資料	101~108	
	內政部公告土地現值及公告地價占一般正常交易價格百分比	101~108	
	網蒐二手車價格資料	110	
上市櫃公司股價資料	100~108		

資料來源：作者自行整理。

原車輛財產檔未包含財產價值資訊，故另透過網蒐二手車交易資訊，以廠牌、排氣數、出廠年份為特徵值，建立預測車價之線性迴歸方程式，估計車輛財產之市價。投資財產則透過上市櫃公司股價資料，將投資人持有之上市櫃公司股票還原為市價，估算出投資財產之市價。

(二) 各類財產地址區位化

基於房屋所在地生活機能可能會影響其出租與否，本主題試圖將各房屋稅稅籍所處之地區，建立相關之區位特徵。理論上來說，最佳的做法應以各房屋稅籍所在地點為中心，計算各生活機能條件之距離，惟基於個資保護原則，現行作業環境並無提供原始地址欄位。取而代之，本主題以村里為最小區域單位，建立生活機能特徵值。故須將外蒐之地址資訊轉換為村里標記，並將外蒐生活機能資料轉換為村里特徵資訊。

肆、資料探索性分析

一、使用特徵

考量影響房屋是否出租之因素多元，若僅以過往財政部所掌握特徵加以建模試行，恐無法突破窠臼。故本主題蒐之各類特徵資料，依其類型可分為三方面，(一) 房屋(建物) 相關特徵，如營業用面積、住家用面積、折舊年度等；

(二) 房屋坐落區位特徵，如坐落村里內房屋、教育及醫療機構數量等；(三) 房屋持有人特徵，如持有房屋數量、持有房屋數量占該村里房屋數比重、所得收入及財產金額、年齡、家庭結構等。

二、資料標註方法

為建立偵測房屋租賃所得模型，須先挑選核定具有租賃房屋所得房屋，作為模型訓練標的。透過綜合所得稅核定所得檔，挑選符合之所得格式代號及註記³，並將相對應之所得資料與房屋稅稅籍檔連結。將有租賃行為之房屋稅籍編號標

記為 1，餘則標記為 0。

三、資料探索成果

(一) 現行租賃所得資料侷限於特定來源

觀察綜所稅核定所得檔，自然人之間房屋租賃交易於現行財政部所蒐集資料中屬相對少數者。108 年租賃所得來源中，租客(所得給付人) 身分僅 1 成屬自然人。其中，若逾 9 成屬非扣繳資料，亦即代表多數納稅義務人非屬主動申報。再者，其非扣繳資料來源，主要來自房屋租金補貼、法院公證案件及透過營業稅交叉查核所提供之房屋租賃所得案件，三者皆非常見之房屋租賃態樣。

(二) 現行租賃所得資料與外蒐資料有顯著差異

若以現行租賃所得資料與外蒐資料比較，其在地區、屋齡及坪數皆有顯著差異。地區方面，前五大地區件數分布僅兩個地區與外蒐資料一致，其餘皆不同；屋齡方面，財政部租賃所得資料屋

齡明顯較高，實價登錄則較低；房屋坪數方面，財政部租賃所得資料坪數相對集中於 30 坪左右，591 租屋平臺則因可能多屬於雅房出租，坪數高峰集中於 10 至 15 坪間（圖 1）。

（三）標記資料極度不平衡

由於目前具房屋租賃所

得資料量相對較少，導致標記為 1 者（有出租者）與為 0 者之間存在高度不平衡。以 103 年至 108 年資料觀察（圖 2），出租標記資料之占比僅有整體 1%。按縣市別觀察，亦存在頗大差異，六都等人口相對密集及工作機會較多之縣市，其出租標

記比重皆在 1% 以上，但澎湖縣、臺東縣等則占比不及 0.4%。面對資料不平衡之狀況，後續在建立模型之成效指標選取上，須更為謹慎，以避免因擇取指標偏誤，導致模型效力詮釋上之誤用。

伍、資料分析策略

一、分析架構及資料訓練範圍

本主題採監督式學習，變數可區分為房屋特徵、坐落區位特徵及屋主特徵等三個面向（下頁圖 3）。模型採用逐年訓練及驗證方式進行，以當年度資料訓練、隔年資料驗證，來觀察模型於歷史資料中的預測效力。訓練資料年度為 103 年至 107 年，相關資料訓練筆數及完成標記之資料筆數參考圖 2。

二、分析方法、驗證機制及成效指標

（一）分析方法

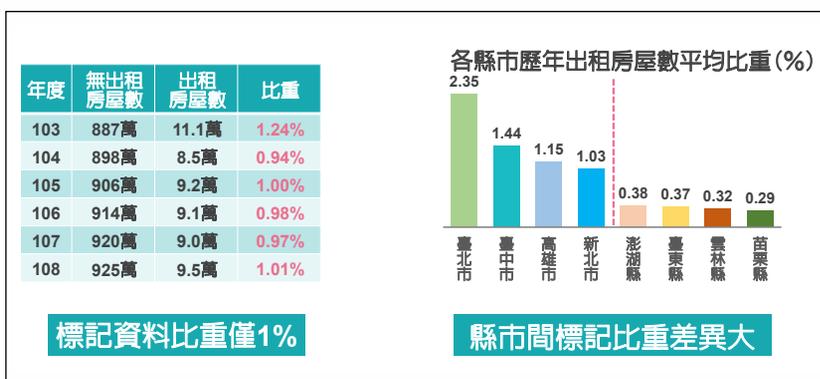
本主題屬預測二元分類議題，故採用分類效果及

圖 1 現行租賃所得資料與外蒐資料差異



資料來源：作者自行繪製。

圖 2 標記資料不平衡



資料來源：作者自行繪製。

論述 》 統計 · 調查

處理資料效率高，常見於 Kaggle 等相關競賽中使用的 LightGBM 演算法 (Light Gradient Boosting Machine，輕量化梯度提升機)。

(二) 驗證機制

因出租與未出租房屋之比例過於懸殊，本主題係屬不平衡資料 (imbalanced data)。而為確保模型於訓練時能充分反映資料特徵，

採用常見的 K 折交叉驗證 (k-fold cross-validation)，來避免模型訓練結果出現過度配適 (overfitting) 或選擇偏差 (selection bias)。本主題設定 k 為 5，將訓練集資料拆分為 5 個子集合進行交叉驗證。

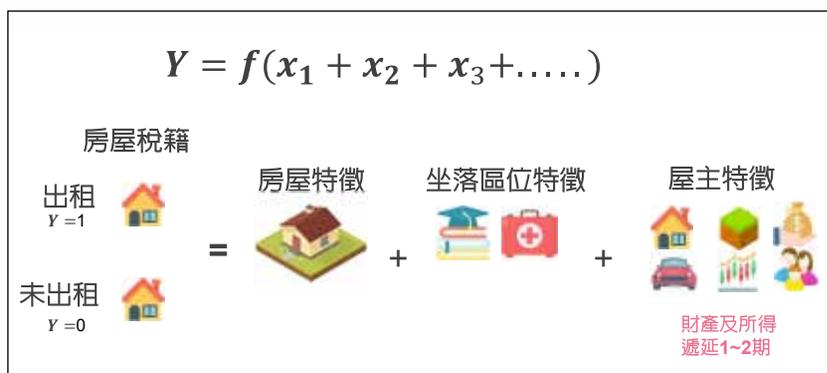
(三) 成效指標

常見的機器學習預測效力指標包含精確率 (Precision)⁴

、召回率 (Recall)⁵、準確度 (Accuracy)、接收者操作特徵曲線下面積 (Receiver Operating Characteristic curve Area Under Curve，簡稱 ROC AUC)、F1 分數 (F1 score) 等指標，其計算方式與適用的狀況各有所異。

茲以圖 4 說明，如屬模型未預測到的後果影響較大，建議使用 Recall 指標，因其計算的是 TP/(TP+FN) 占比，著重於模型預測結果所能掌握到實際陽性的比重，適用於產線的瑕疵品偵測、流感疫苗接種等議題。若屬模型預測錯誤的後果影響較大，建議使用 Precision 指標，因其計算的是 TP/(TP+FP)，著重於模型預測為陽性當中，確實屬於陽性的比重，適用於疾病測試、驗孕測試或發單課稅等議題。其他如 Accuracy、ROC AUC，因其計算方式會同時考量 TP、TN，易受到資料不平衡影響而誤判預測效力，故不建議使用。

圖 3 模型概念說明



資料來源：作者自行繪製。

圖 4 衡量模型效力指標考量因素



資料來源：作者自行繪製。

本主題欲預測房屋之出租行為，並藉以輔助課稅使用，故建議採 Precision 指標衡量預測效力。另考量因資料存在高度不平衡現象，僅約 1% 屬於有註記租屋者（Positive，陽性）。若直接採用 Precision 衡量模型效力，亦有所失當。故採用實務上常用於推薦系統時使用的 Precision at K，作為觀

察預測成效之指標。

所謂的 Precision at K 係將模型預測結果按照機率之高至低排序，並設定 K 值，計算前 K 筆資料當中，預測正確之比率。如 K 值為 10，代表預測為陽性的機率最高之 10 筆資料中，確實為陽性的比重（圖 5）。

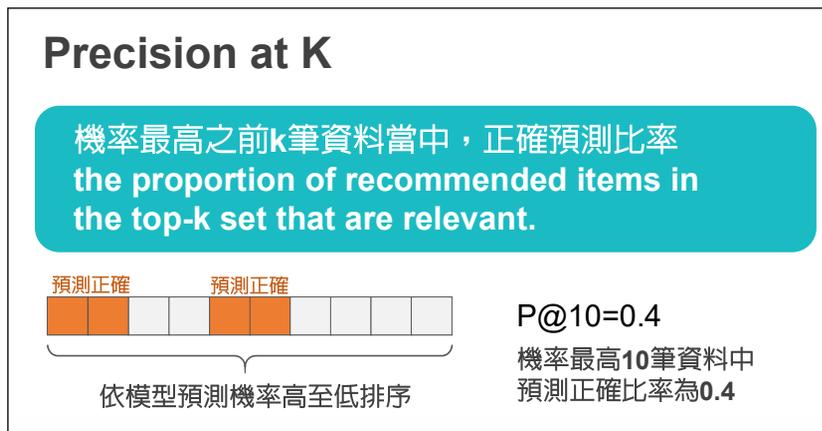
陸、模型效益評估

以常見機器學習模型效能評估指標觀察，準確度（Accuracy）與接收者操作特徵曲線下面積（ROC AUC）皆在 9 成以上，但其係因出租與非出租房屋標記數量懸殊所導致，故此兩項指標並無參考價值。觀察各年度精確率（Precision）大致介於 5 至 7 成間，召回率（Recall）則介於 3 至 4 成間，綜合性 F1 指標落在 5 成以下。顯示在資料不平衡的現象下，模型預測的效力表現不如預期（圖 6）。

若改以 Precision at K 為效力指標，觀察模型預測機率最高之前 1 千筆資料，精確率僅 105 年因涉及隔年稅制調整⁶而降低，其餘各年則介於 0.689 至 0.798 間，預測效力大幅提升。若以臺北市、新北市及臺中市為資料來源，於限縮縣市下建模預測，精確率可進一步提高，各年介於 0.762 至 0.846 間，顯示本模型確有其潛在實務效益。

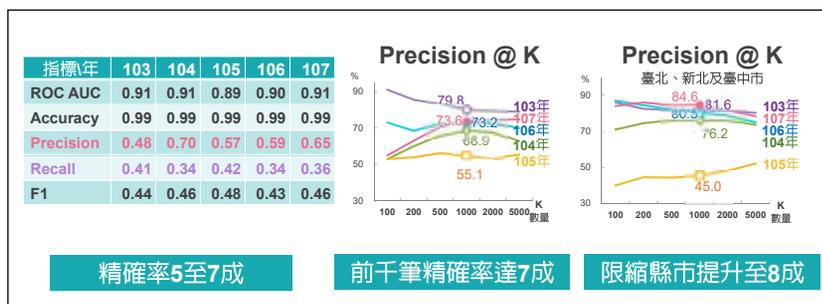
統整各年度模型驗證測試結果之前 20 大重要特徵值，確實如建模時所預估，三大面

圖 5 Precision at K 指標



資料來源：作者自行繪製。

圖 6 各年度模型解釋效力值



資料來源：作者自行繪製。

