

1132 - 機器學習與其在行為科學之應用

## 預測網路沉癮傾向與其解釋性

111304019 統計三 林承佑

111304011 統計三 方品皓

111304015 統計三 陳則銓

## 目錄

### 一、INTRODUCTION

1. 研究背景
2. 資料來源
3. 關注主題

### 二、METHOD

1. 資料介紹
  - (1) 心裡健康因子
  - (2) 後疫情生活型態
  - (3) 人口背景
  - (4) 網路沉迷指標
  - (5) 備註
2. 建模策略
  - (1) 資料不平衡處理
  - (2) 模型內建權重調整:
  - (3) 優化決策閾值 (*Optimizing Decision Threshold*):
3. 模型選擇
4. 資料切分、訓練、驗證、調參
5. 評估指標
  - (1) 評估指標的選擇:
  - (2) *Precision-Recall* 曲線 (*PR Curve*):

### 三、RESULTS

1. 分析結果
  - (1) *Lasso* 特徵篩選:
  - (2) 單一模型表現:
  - (3) 集成模型優勢:
2. 模型成效
3. 特徵重要性
4. 檢定結果

### 四、CONCLUSIONS & DISCUSSION

## 一、Introduction

### 1. 研究背景

在後疫情時代，數位生活已成為常態，與此同時，大眾對於網路沉迷（Internet Addiction）問題的擔憂也日益加劇。此現象不僅影響個人的心理健康、學業與職業表現，也可能引發家庭與社會問題。因此，若能及早識別出具有網路沉迷傾向的高風險族群，並提供適當的輔導與介入，將是公共衛生領域一項重要的課題。

### 2. 資料來源

為「數位發展部」主辦、由「聯合行銷研究股份有限公司」執行之《111 年網路沉迷研究調查》，目前由中央研究院社會科學資料庫彙整與發布（計畫編號 AE20004）。

### 3. 關注主題

「能否透過一個人的心理狀態、疫情經歷與人口背景特徵，準確預測其是否有網路沉迷傾向（非量表）？」可否藉由人口特徵與心理指標，預測網路沉迷高風險族群，協助早期識別與介入。

我們期待找出網路沉迷傾向的關鍵預測因子，可否藉由人口特徵與心理指標，預測網路沉迷高風險族群，並提供適當的輔導與介入。

## 二、Method

### 1. 資料介紹

包含人口學特徵、疫情相關經歷、心理狀態指標以及網路使用行為。目標變數 target 是二元的 (0: 無成癮傾向, 1: 有成癮傾向)，且存在嚴重的類別不平衡 (1356 vs. 130)，這將是我們建模時需要重點處理的問題。

#### (1) 心裡健康因子

- 包含：生活無聊感、課業或工作壓力、難以集中精神。

#### (2) 後疫情生活型態

- 包含：因疫情失去工作或工作減量、擔心確診減少外出、或居家上課上班的情況等；海報上列出幾個我們「先」認為較為重要的類別，那其中還有其他許多項目。

#### (3) 人口背景

- 包含：年齡（保留所以有原本的分層）、職業（原本的太細節，少數職業樣本過低，所以將定義和標準稍微放寬，再做一次分類）、性別、教育程度、地區（這裡額外再分成直轄市多一個特徵）

#### (4) 網路沉迷指標

- 網路沉迷指標，CIAS-10 總分是否  $\geq 28$  作為分類項目，這是台大的「陳氏網路沈迷量表」，如果在這十題當中，有任一題選擇不願透露，導致分數計算可能會有偏誤的資料，都選擇刪除。（並不多）（有無網路成癮的狀況其實非常不平衡，大約 9:1 甚至更少）

#### (5) 備註

- 「近六個月上網時間」的 HH:MM 格式轉換為數值（中位數填補，極少有缺失值）；有序類別（如非常同意到非常不同意）轉為 Ordinal Encoding；無序類別則用 One-Hot Encoding。

### 2. 建模策略

### (1) 資料不平衡處理

分層抽樣 (Stratified Sampling): 在將資料切分為訓練集與測試集時，使用 `stratify=y` 的參數，確保訓練集與測試集中的目標變數分佈與原始資料集一致。

### (2) 模型內建權重調整:

XGBoost: 在建立 XGBoost 分類器時，透過計算 `scale_pos_weight` 參數（反類別頻率 0 類的數量 / 1 類的數量）來加大對少數類別（有成癮傾向）預測錯誤的懲罰。

隨機森林 (Random Forest): 在建立隨機森林分類器時，直接設定 `class_weight='balanced'` 參數，讓模型自動根據類別頻率的反比來調整權重。

### (3) 優化決策閾值 (Optimizing Decision Threshold):

訓練完集成模型後，專案並未使用預設的 0.5 作為判斷成癮與否的機率分界點。而是透過繪製 Precision-Recall 曲線，找到了一個能讓 F1-Score 最大化的最佳閾值（0.5825）。其兼顧 Precision 和 Recall：

- Precision 關注於，在所有「預測」為沉迷者，真的為沉迷者的情況。
- Recall 則關注於，在所有「真正」為沉迷者中，預測正確的情況。

因此，我們希望 Recall 越高越好，這樣就可以正確預測；但我們也同時希望 Precision 越高越好，才不會都一直判斷錯誤，把不是沉迷者的結果都判斷為沉迷者，這樣就會太多人被誤判。從而顯著提升了模型在少數類別上的綜合表現（F1-Score 從 0.39 提升至 0.45）。

## 3. 模型選擇

- 邏輯斯迴歸 (Logistic Regression): 作為基準模型 (Baseline)。
- 隨機森林 (Random Forest): 是基於 Bagging 的集成學習模型，透過建立多棵決策樹並投票，有效降低過擬合的情況。
- XGBoost：是基於提升法 Boosting 的集成學習模型，以其高效能和高準確率著稱，是業界和競賽中非常主流的模式。

投票分類器 (Voting Classifier): 為了整合上述三個模型的優點，專案最終建立了一個軟投票 (soft voting) 的集成模型。此模型會根據各模型在交叉驗證中的 F1 分

數給予不同權重，綜合它們的預測機率來做出最終判斷，通常能獲得比任何單一模型都更穩健的表現。

#### 4. 資料切分、訓練、驗證、調參

- 75% 訓練集和 25% 測試集
- 採用分層抽樣確保類別比例一致
- RepeatedStratifiedKFold (10 折交叉驗證，重複 3 次)
  - 比單純的 K-Fold 驗證更為可靠，它保持了每折中類別比例的平衡，還透過多次重複降低了由隨機切分帶來的結果波動。
- 使用 GridSearchCV 進行全面的網格搜索，以找到在 F1-score 指標下表現最佳的超參數組合。

#### 5. 評估指標

##### (1) 評估指標的選擇:

專案選擇 F1-score 作為模型調參與評估的主要指標，因為它能同時兼顧精確率 (Precision) 和召回率 (Recall)，比準確率 (Accuracy) 更能反映模型在不平衡資料上的真實表現

##### (2) Precision-Recall 曲線 (PR Curve):

在集成模型建立後，專案繪製了 PR 曲線，並計算了曲線下面積 (AUC)。更重要的是，透過這條曲線找到了能讓 F1-score 最大化的最佳預測機率閾值 (threshold)，而非使用預設的 0.5

F1-Score: 這是本次專案最核心的評估指標，用於超參數搜索和最終模型比較。它綜合了精確率和召回率，特別適用於不平衡資料。Precision (精確率) 與 Recall (召回率): 用於評估模型對於正類別 (有成癮傾向) 的預測表現。精確率代表「模型預測為正的樣本中，有多少是真的正樣本」；召回率代表「所有真的正樣本中，有多少被模型成功預測出來」。

### 三、Results

#### 1. 分析結果

##### (1) Lasso 特徵篩選:

在模型建立初期，Lasso 迴歸篩選出了 16 個重要特徵。結果顯示，「難以集中精神」是風險最高的正相關特徵，而「無法節制時間的網路活動項目\_其他\_1」是保護性最強的負相關特徵。但事實上，做過特徵篩選的結果與不做的結果相比，並沒有顯著地提升指標。因此接下來的表現都沒有使用特徵篩選，將所有資料都丟入模型當中進行訓練。

##### (2) 單一模型表現:

在三個獨立模型中，XGBoost 在交叉驗證和測試集上的 F1-score 表現都是最好的，顯示了其處理這類複雜資料的強大能力。邏輯斯迴歸和 XGBoost 都達到了 70% 的召回率，表示它們成功識別了七成的成癮傾向者，但代價是精確率較低（約 23%-27%），產生了較多的誤判。

##### (3) 集成模型優勢:

最值得注意的是，加權投票的集成模型，在經過 PR 曲線優化預測閾值後，取得了最優的綜合表現。雖然召回率略有下降（從 70% 降至 58%），但精確率大幅提升至 37%，最終使 F1-score 顯著提高到 0.45。這意味著最終模型在「找到真正成癮者」和「避免將非成癮者誤判」之間取得了更好的平衡，同時整體準確率也達到了最高的 88%。

整合成一個投票分類器 (Voting Classifier)。我們將採用「軟投票 (Soft Voting)」，它會對三個模型的預測機率進行加權平均，通常比「硬投票」（直接取決於票數）效果更佳。根據 PR 曲線上 F1-score 的最大值來確定最佳分類閾值。

#### 2. 模型成效

### 模型成效

註：紅字代表效能最佳

Threshold: 0.5825

評估指標	Log Regression	Random Forest	XGBoost	Esemble Models
Accuracy	0.77	0.83	0.81	0.88
Recall	0.70	0.55	0.70	0.58
Precision	0.23	0.26	0.27	0.37
F1-Score	0.35	0.36	0.39	0.45

### 3. 特徵重要性

- Y 軸 (特徵名稱): 列出了模型使用的所有特徵。
- X 軸 (F1 分數下降值): 代表打亂該特徵後，模型 F1-Score 下降的幅度。X 值越大，代表該特徵越重要。
- 圖形 (箱型圖): 由於隨機打亂的過程會重複多次（在這個專案中重複了 10 次），箱型圖顯示了這 10 次分數下降值的分布情況。

最重要的特徵：「無法節制時間的網路活動\_其他」是模型眼中最重要的特徵，打亂它會導致 F1-Score 平均下降約 0.07。其次是「難以集中精神」、「生活無聊感」和「擔心確診情況」。

中等重要的特徵：「近六個月上網時間\_分鐘」、「無法節制時間的網路活動\_看網路社群」等特徵也對模型有顯著的貢獻。

較不重要的特徵：位於圖表最下方的特徵，如「性別」、「教育程度」、「是否居住直轄市」等，它們的 F1 分數下降值接近於 0，代表在這個模型中，它們對預測網路成癮傾向的貢獻非常小。

比單純的重要性排序提供了更豐富的資訊，SHAP 揭示了特徵的數值大小如何影響模型的預測方向和強度。

- Y 軸 (特徵列表): 特徵按照全局重要性排序，最重要的在最上面（這與 SHAP Summary Plot 的結論一致）。
- X 軸 (SHAP Value): 代表特徵對模型輸出的影響力。
- 正值 ( $> 0$ ): 表示該特徵的數值將模型的預測推向「有成癮傾向」。
- 負值 ( $< 0$ ): 表示該特徵的數值將模型的預測推向「無成癮傾向」。
- 顏色 (Feature Value): 每個點代表測試集中的一個樣本。點的顏色代表該特



徵在該樣本中的原始數值大小。紅色代表數值高，藍色代表數值低。

A. 無法節制時間的網路活動\_其他\_0：

- I. 分析：這是一個分類特徵。紅色的點（代表數值為 1，即「有」這項活動）幾乎全部集中在 X 軸的正半側，具有很高的正 SHAP 值。藍色的點（代表數值為 0，即「沒有」這項活動）則集中在負半側。
- II. 結論：如果一個人在「其他」項目上無法節制上網時間，會強烈地增加其被預測為網路成癮的機率。

B. 難以集中精神：

- I. 分析：紅色的點（代表「難以集中精神」的程度高）普遍有正的 SHAP 值，而藍色的點（程度低）則有負的 SHAP 值。
- II. 結論：「難以集中精神」的程度越高，被模型判定為有成癮傾向的風險也越高。這是一個清晰的正相關關係。

C. 近六個月上網時間\_分鐘：

- I. 分析：同樣地，紅色的點（上網時間長）傾向於有正的 SHAP 值，藍色的點（上網時間短）傾向於有負的 SHAP 值。
- II. 結論：上網時間越長，網路成癮的風險越高。

#### 4. 檢定結果

統計檢定驗證了核心特徵的顯著性，其為我們的發現提供了另一層次的證據。這個檢定可以看作是對機器學習模型發現的關聯性進行統計顯著性的確認。

- ord\_\_難以集中精神 的 p-value 為 0.028 ( $< 0.05$ )，達到統計顯著水平。這證實了「難以集中精神」不僅是模型的重要預測因子，其與成癮傾向之間的正向關聯（estimate = 0.0118）在統計上也是成立的。
- ord\_\_疫情後上網狀況 的 p-value 為 0.015 ( $< 0.05$ )，同樣達到統計顯著水平。這也證實了疫情後上網情況的改變，是預測成癮傾向的一個統計上顯著的因素。

#### 四、Conclusions & Discussion

本專案的核心精神，跳脫機器學習中對「準確率」的迷思。因為我們的資料存在不平衡的問題，而高準確率往往掩蓋了模型對少數關鍵個案（有成癮傾向者）的無知。只要全部都猜是沉迷，準確率的計算就會超高！

依照結果發現，加權集成模型在綜合評估指標上取得了較佳的結果。更重要的是，在解釋重要變數中，主要影響網路沉迷傾向的關鍵因子。

起初我們可能會認為，年齡、或居住的縣市會是很重要的影響因素；然而事實上，心理健康狀態與特定網路行為是更強的預測變數，像是「難以集中精神」、「生活無聊感」以及「無法節制地從事其他網路活動」等特徵尤為重要。後續統計檢定，亦證實了「難以集中精神」與「疫情後上網狀況」的關聯性具備統計學上的顯著性

我們不再滿足於一個能正確分類大多數「無成癮傾向」樣本的模型，而是從「忽略錯誤」到「主動發現」，透過審視不同的評估指標（如 F1-Score）與混淆矩陣，主動去探尋那些被成功識別出來的少數個案。這正是早期識別與介入策略的核心所在——其價值不在於完美地將所有人分類，而在於精準地找到最需要幫助的那些人。