巨量資料分析課程報告書

偷電偵測

組員：陳雅柔、劉家維、邱奕銓、
羅然莉、楊家瑋、張子恩、韓明澄

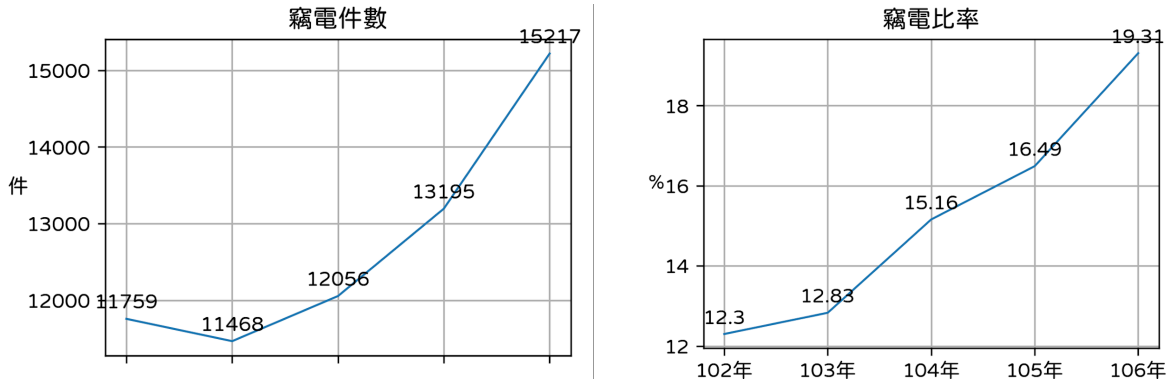
目錄

1 簡介

本研究旨在找尋有效偵測竊電的方法，並且以這些方法製作應用程式以期達到協助判斷竊電的可能性，我們系統地評估了現有的檢測方法，最終選擇了三種模型，寬深 CNN、CLOF、CBLOF 做為判斷依據，他們各自適合不同的分析情境，結果顯示，新的檢測方法在防止電力盜竊方面具有潛力。

1.1 研究動機

全球每年因電力盜竊造成的經濟損失高達近千億美元。電力盜竊問題不僅對經濟造成巨大損失，還可能影響電力系統的穩定性和安全性。在台灣，電力盜竊問題同樣嚴重且逐年加劇。台電公司 106 年度全系統發購電電量達 2,310.80 億度，扣除抽蓄用電、公司自用電及售電量後，全年線路損失量為 88.27 億度，線路損失率為 3.82% 。102 年度至 106 年度查獲竊電情況如圖 1.1，反映電力被竊用的情形有日趨嚴重之勢。



(a) 竊電件數　　(b) 竊電比例

圖 1.1: 竊電情況

現行偵測竊電的研究大致分為四個方向，賽局理論、電網分析、硬體分析、機器學習，由於機器學習以外都有相當程度的限制，我們主要探討機器學習相關的方式。

1.2 專案背景與目的

本研究的背景在於應對現代化電力系統中日益嚴重的電力盜竊問題。隨著篡改電表讀數的方法變得更加多樣和隱秘，現有的檢測方法面臨新的挑戰。本研究的主要目的是基於 SGCC(STATE GRID Corporation of China) 資料集，系統地調查和評估各種電力盜竊檢測方法。包括基於機器學習和測量不匹配的方法。通過這些分析，我們期望能夠提出檢測方法，以方便我們利用應用程式檢測竊電者，以達到追回欠繳電費的目的。

2 資料分析與方法

2.1 資料觀察

本研究所使用的資料集涵蓋了 42372 個電力用戶在 1035 天內的用電數據，資料記錄期間從 2014 年 1 月 1 日至 2016 年 10 月 31 日，以下是資料集的主要特徵。

- 日期格式 (MM/DD/YYYY)：每筆記錄表示該日的電力消耗量。
- 電網編號 (CONSNO)：電網編號為字符串類型，代表每個電網的唯一身份。
- 標誌欄位 (FLAG)：此欄位用於標示是否存在竊電行為，其中 0 表示無竊電，1 表示有竊電。在此資料集中，標示為 0 的記錄有 38637 筆，而標示為 1 的記錄則有 3585 筆，如圖 2.1 。

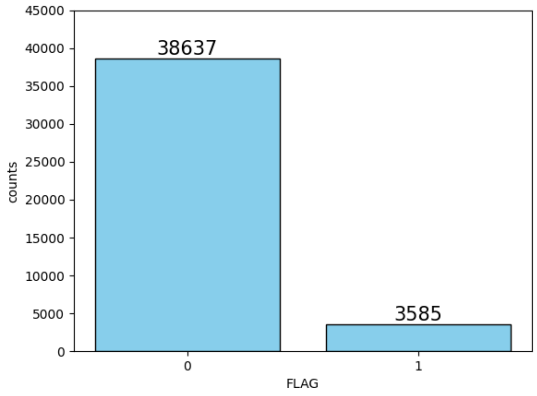


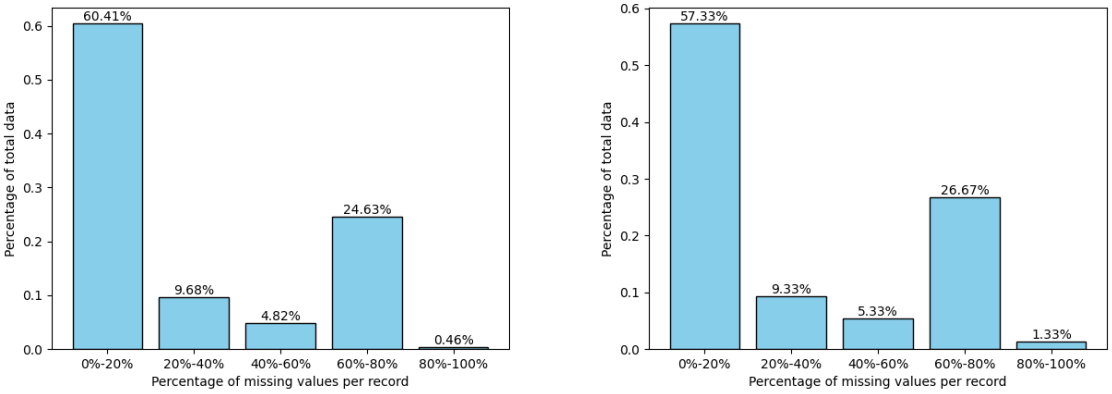
圖 2.1: 訓練資料中有無偷電數量比較圖

2.1.1 資料問題

- 資料缺失值與離群值多，若不處理將會影響後續的分析及預測，個數統計如表 2.1 ，百分比見圖 2.2

資料集	訓練集資料	測試集資料	總個數
缺失值個數	11174212	42872	11217084
離群值個數	1705795	6105	1711900

表 2.1: 缺失值與離群值統計



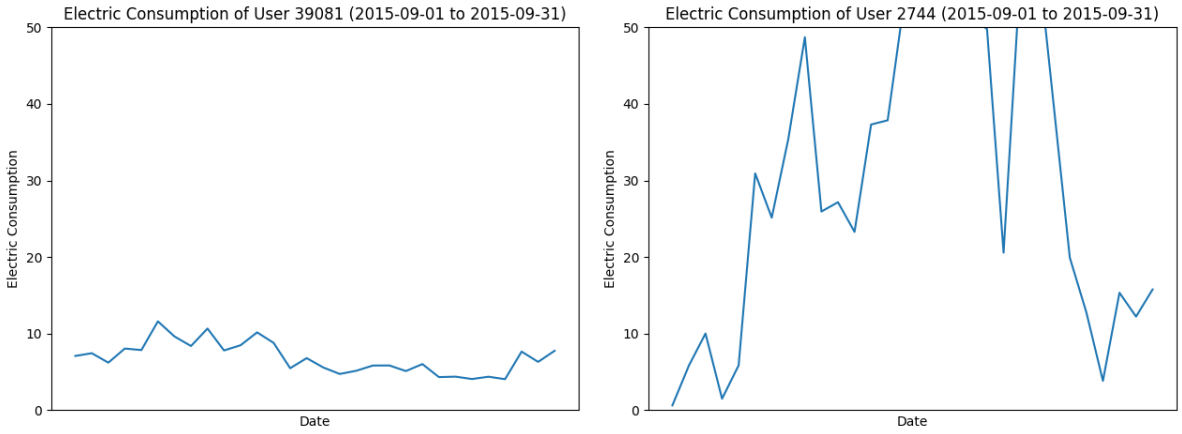
(a) 訓練集　　(b) 測試集

圖 2.2: 每個電網的缺失值百分比

- 資料極度不平衡資料，若不處理，可能會導致準確度高，但精確度低的問題。
- 不同用戶的用電量差異很大，需要進行資料標準化或正規化處理，以提高模型的效果。
- 除了缺失值，即便是正常用戶，可能因為出國或其他原因，導致用電量在幾個月內為零，這也會對預測結果造成影響。
- 資料集的日期並未照順序排序，在視覺化或是分析可能會遇到問題，因此將資料按時間排序也是前處理的重點。

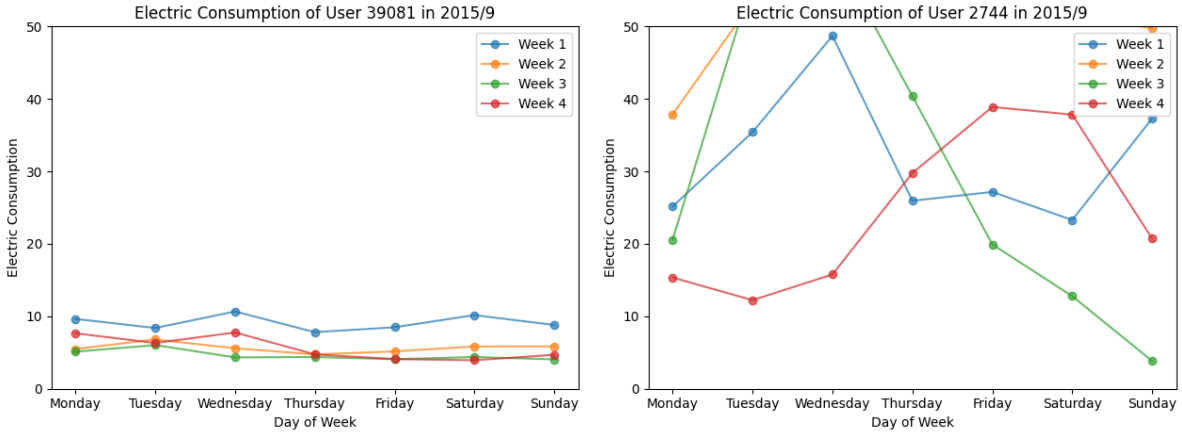
2.1.2 資料視覺化

將用電量分別畫出了正常用戶和偷電用戶在一年中每日用電量如圖 2.3 (a),(b)、單月每週用電量見圖 2.4 (a),(b) 、以及按月份比較一年中的每日用電量如圖 2.5 (a),(b) 。觀察結果顯示，竊電者的用電起伏較大，而正常用電者的用電則較為具有週期性。基於這一發現，在處理缺失值時，我們除了使用常見的前後平均值法外，還採用了每個月該星期的平均值來補全缺失數據，然而仍有許多差異是肉眼難以辨識的，因此需要透過模型訓練、預測和排序來精確識別。



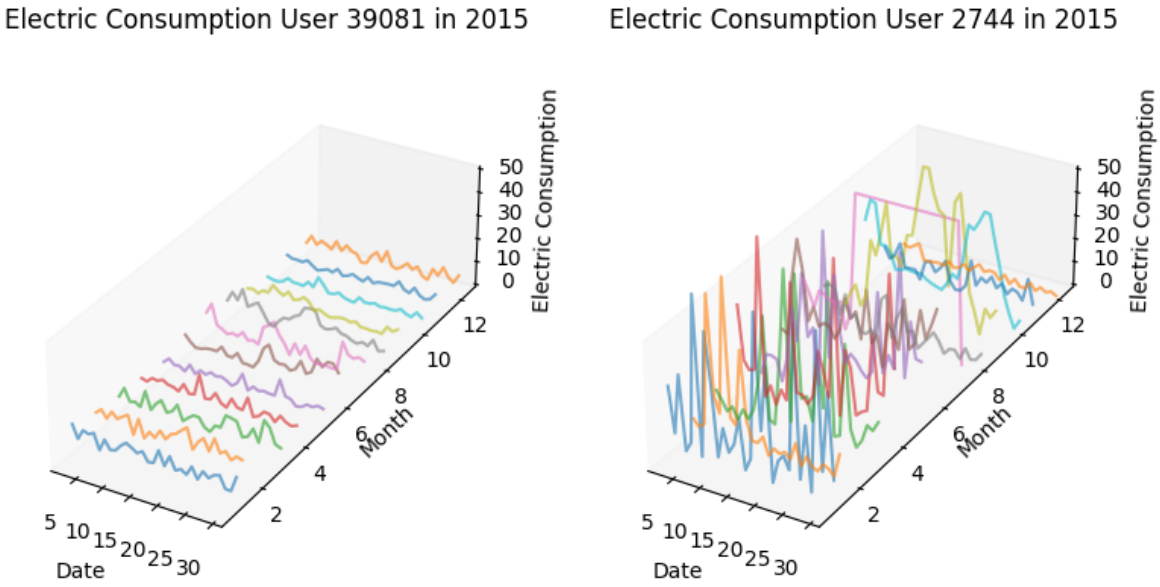
(a) 正常用戶 (b) 偷電用戶

圖 2.3: 2015 年 9 月日用電量



(a) 正常用戶 (b) 偷電用戶

圖 2.4: 2015 年 9 月週用電量



(a) 正常用戶 (b) 偷電用戶

圖 2.5: 2015 年的月用電量

2.2 資料前處理與分析方法原理

2.2.1 寬深 CNN(Wind and Deep CNN)

資料前處理

1. 修正資料排序：因為此為時間序列資料，若時間沒有按照正常順序排序的話會很大程度的影響模型判斷。因此先將原先雜亂的日期整理成按照時間先後排序的資料。

2. 修正離群值：如果某個資料點超過均值的三個標準差，則認為該點是錯誤的，使用 Three-sigma rule of thumb 來修正，如式 (2.1) 。

$$f(x_i) = \begin{cases} \text{avg}(x) + 2 * \text{std}(x) & \text{if } x_i > \text{avg}(x) + 2 * \text{std}(x) \\ x_i & \text{otherwise} \end{cases} \tag{2.1}$$

3. 處理缺失值：當該筆中有資料缺失時，使用該筆資料中的其他可用資料均值來替換缺失值，見式 (2.2) 。

$$f(x_i) = \begin{cases} \frac{x_{i-1}+x_{i+1}}{2} & \text{if } x_i \in \text{NaN}, x_{i-1}, x_{i+1} \notin \text{NaN} \\ 0 & \text{if } x_i \in \text{NaN}, x_{i-1} \text{ or } x_{i+1} \in \text{NaN} \\ x_i & \text{if } x_i \notin \text{NaN} \end{cases} \tag{2.2}$$

4. 正規化：如式 (2.3)。

$$f(x_i) = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{2.3}$$

分析方法原理

根據「資料視覺化的發現」，得知了能從是否具有週期性來去判斷是否有偷電。為了能有效的應用這個發現，使用了寬深 CNN 來去分析電力消費數據。

寬深 CNN 的框架主要由寬組件 (Wind Component) 跟深度卷積神經網絡組件 (CNN Component) 兩個部分組成，寬組件可以學習全局知識，而深度卷積神經網絡組可以捕捉電力消費數據的週期性。這個模型整合了寬組件和深 CNN 組件的優點，因此在電力盜竊檢測中能有優秀的表現，其模型的運作方式如圖 2.6 所示。

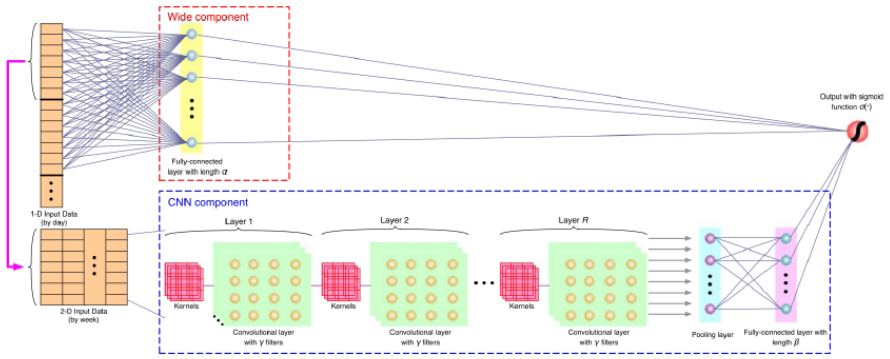


圖 2.6: 訓練資料中有無偷電數量比較圖

寬深卷積神經網絡的整合通過整合這兩部分的輸出，最終形成一個具有強大特徵學習能力的模型。

- 寬組件 (Wind Component) 寬組件是一個全連接層，主要用於從一維電力消費數據中學習全局知識。電力消費數據通常是隨時間變化的時間序列數據，而這些數據可能顯示出周期性模式或非周期性模式。寬組件的設計目的是捕捉這些全局模式。全連接層中的每個神經元根據式 (2.4) 使用一維電力消費數據計算其自身的得分：

$$y_j := \sum_{i=1}^{n} w_{i,j} x_i + b_1 \tag{2.4}$$

其中 y_j 是第 j 個神經元在全連接層的輸出，n 是一維輸入數據 x 的長度，$w\{i,j\}$ 表示第 i 個輸入值和第 j 個神經元之間的權重，b_1 是偏差。計算後，它將通過激活函數將此值發送到更高層的連接單元，以確定它對下一步預測的貢獻程度。激活函數如式 (2.5) 所示：

$$u_j := f(y_j) = \max(0, y_j) \tag{2.5}$$

- 深度卷積神經網絡組件 (CNN Component)
  - 卷積層：每個卷積層包含多個卷積過濾器，這些過濾器專門設計來捕捉電力消費數據中的特定模式，如周期性或非周期性模式。卷積過程將輸入數據轉換為特徵圖，通過多個卷積層的處理，逐漸提取更高層次的特徵 。

- 技術細節：深度 CNN 組件的訓練過程中，輸入的二維電力消費數據將通過多個卷積層，這些層能夠有效地捕捉二維數據的特徵。在實驗中選擇了不同數量的卷積層（例如，y 個過濾器），以調整模型性能 。

分析與驗證結果

CNN 模型驗證指標，見圖 2.7 所示。

- 不錯的區分能力 (AUC: 0.81)：模型的 AUC 值為 0.81 ，顯示出它有較好的整體分類性能，能夠有效區分大部分的盜竊行為和正常行為。
- 中等的 F1 分數 (F1 Score: 0.39)：儘管模型在高優先級位置上的精確度很高，但整體的精確度和召回率之間的權衡不夠理想，這表明模型在全面檢測所有盜竊行為時可能有一定的不足。
- 不錯的前 100 名和前 200 名預測精確率 (MAP@100, MAP@200: 0.91,0.9)：這兩個指標表明，模型在前 100 和 200 個排序位置上的平均精確度非常高，能夠有效地將最可能的盜竊行為排在前面，這對於實際應用中需要優先處理的場景非常有用。

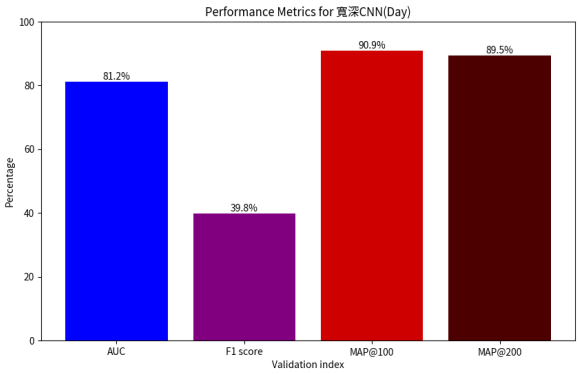


圖 2.7: CNN 模型指標

改進方向

- 提高召回率: 通過調整模型或引入更多特徵來提高召回率，從而提升 F1 Score。

- 平衡精確率和召回率：考慮使用其他方法，如調整閾值或引入代價敏感學習，以平衡精確率和召回率，如圖 2.8 。
- 進一步分析特徵：對於模型的特徵進行進一步分析，找出哪些特徵對於檢測盜竊行為更為關鍵，並對模型進行優化。

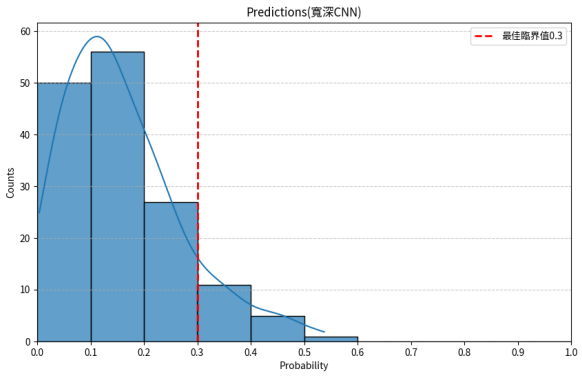


圖 2.8: CNN 模型預測結果與臨界值

2.2.2 Cluster-based Local Outlier Factor(CLOF)

資料前處理

1. 處理缺失資料：當負載向量中的資料缺失時，使用該向量中其他可用資料的均值來替換缺失值，如式 (2.6) 。

$$G(\tilde{u}_{i,t}) = \begin{cases} \text{mean}(\tilde{u}_i) & \text{if } \tilde{u}_{i,t} \in \text{NaN} \\ \tilde{u}_{i,t} & \text{otherwise} \end{cases} \tag{2.6}$$

2. 修正錯誤資料：如果某個資料點超過均值的三個標準差，則認為該點是錯誤的，使用” 三西格瑪準則” 來修正，如式 (2.7) 。

$$G(\tilde{u}_{i,t}) = \begin{cases} \frac{\tilde{u}_{i,t-1}+\tilde{u}_{i,t+1}}{2} & \text{if } \tilde{u}_{i,t} > 3\sigma(\tilde{u}_i) \text{ and } \tilde{u}_{i,t-1}, \tilde{u}_{i,t+1} \neq \text{NaN} \\ \tilde{u}_{i,t} & \text{otherwise} \end{cases} \tag{2.7}$$

3. 正規化：每個負載向量的每個元素會被該向量的最大值進行歸一化處理。

分析方法原理

CLOF(Cluster-based Local Outlier Factor) 方法將聚類技術與 LOF 方法結合起來，用於識別電力消費數據中的異常行為，特別是電力偷竊。這種方法的主要步驟包括：

1. 聚類分 (Clustering)：
    - 使用 k-means 聚類算法對用戶的電力消費模式進行分組。這一步的目的是根據消費特徵將用戶劃分為若干類別，使得每個類別中的用戶具有相似的消費行為。
    - 對於每個聚類結果，計算聚類中心並衡量每個用戶與其聚類中心的距離。距離較遠的用戶可能是潛在的異常值。
2. 局部離群因子 (Local Outlier Factor, LOF)：
    - 在聚類分析的基礎上，對每個聚類內的用戶應用 LOF 算法。LOF 算法通過評估每個數據點的局部密度，來衡量該數據點相對於其鄰域內其他數據點的異常程度。
    - 具體來說，LOF 計算每個數據點的局部可達密度 (Local Reachability Density, LRD) ，並與其鄰域內其他數據點的 LRD 進行比較。如果一個數據點的 LRD 顯著低於其鄰域內的其他點，則該點被認為是潛在的異常值。
3. 異常檢測框架 (Detection Framework)：
    - 最終，CLOF 方法結合聚類分析和 LOF 的結果，識別並標記那些可能存在異常行為的用戶。
    - 該方法能夠有效地處理數據集中的不平衡問題，並能檢測到各種類型的偷電攻擊，從而提高電力偷竊行為的檢測準確性。

LOF(Local Outlier Factor) 是一種基於密度的異常檢測方法，它通過比較每個數據點與其鄰域內其他數據點的密度來評估該數據點的異常程度。LOF 方法的主要步驟包括：

1. 鄰域定義 (Clustering)

- 對於每個數據點 p，找到其 k 個最近鄰居 (k-nearest neighbors)。這些鄰居構成該數據點的鄰域。

2. 可達距離 (Reachability Distance)

   - 對於數據點 p 和其鄰居 o，計算可達距離 reach-dist(k,p,o)。可達距離是 o 和 p 之間的實際距離以及 o 的 k 距離中較大的一個，如式 (2.8) 。

$$\text{reach-dist}(k,p,o) = max(\text{k-distance}(o), \text{dist}(p,o)) \tag{2.8}$$

     - k-distance(o) 是 o 的 第 k 個最近鄰的距離。

3. 局部可達密度 (Local Reachability Density, LRD) 如式 (2.9) 。

$$\text{LRD}(p) = \left(\frac{\sum_{o \in N_k(p)} \text{reach-dist}(k,p,o)}{|N_k(p)|} \right)^{-1} \tag{2.9}$$

   - 計算數據點 p 的局部可達密度，即 p 的鄰居的平均可達距離的倒數。
   - $N_k(p)$ 是 p 的 k 個最近鄰居的集合。

4. 局部離群因子 (Local Outlier Factor, LOF)

$$\text{LRD}(p) = \left(\frac{\sum_{o \in N_k(p)} \frac{\text{LRD}(o)}{\text{LRD}(p)}}{|N_k(p)|} \right)^{-1} \tag{2.10}$$

   - 計算數據點 p 的局部離群因子
   - LOF 值衡量了 p 的局部密度與其鄰居的局部密度的比率。如果 p 的 LOF 值顯著大於 1，則表示 p 是異常點。

分析與驗證結果

CLOF 模型驗證指標，見圖 2.9 所示。

- 一定的區分能力 (Accuracy: 0.87)：模型在整體上能夠比較準確地區分出竊電用戶和非竊電用戶。
- 低精確率 (Precision: 0.10) ：模型預測的竊電用戶中，只有極少數真正是竊電者，這表示模型有很多錯誤的預測結果。

- 低召回率 (Recall: 0.06)：模型僅檢測到極少部分的竊電用戶，大部分的竊電行為沒有被模型捕捉到。這可能意味著模型過於保守，只在非常明顯的情況下才會做出預測。

- 低的 F1 分數 (F1 Score: 0.08) ：表示模型在精確率和召回率之間的平衡較差，需要大幅改進以提升整體性能。

這些指標表明，儘管模型的準確率較高，但在檢測竊電用戶時存在很大的不足。特別是精確率和召回率都非常低，導致 F1 分數也非常低，說明模型在竊電行為的檢測上並不理想，需要進一步優化。

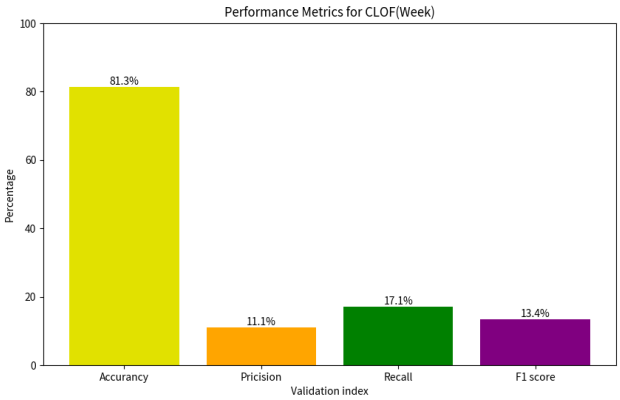


圖 2.9: CLOF 模型指標

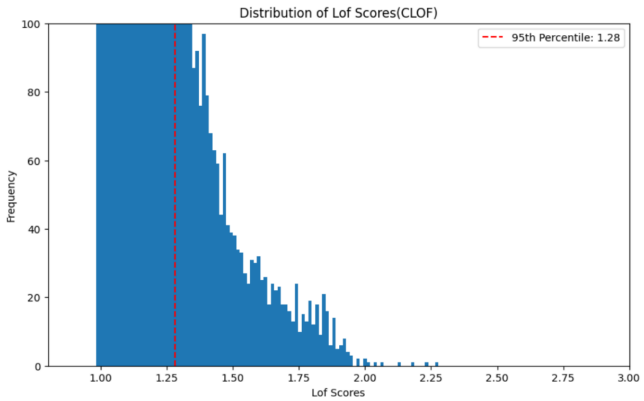


圖 2.10: CLOF 模型預測結果與臨界值

2.2.3 Cluster-Based Local Outlier Factor(CBLOF)

資料前處理

（這裡的前處理均參考《An Ensemble Deep Convolutional Neural Network Model for Electricity Theft Detection in Smart Grids》此篇論文）

1. 處理缺失值：參考了論文提出的填補缺失值方法，考慮了不同季節、月份、工作日和週末的用電模式差異。對於特定天（如星期一）的用電量，使用式 (2.11) 來填補缺失值：

$$f(CM_i) = \begin{cases} \frac{\sum_{n=1}^{x} CM_n}{n} & \text{若} 1 \leq n \leq 4, n \in \mathbb{N} \\ CM_i & \text{若} M_i \notin \text{NaN} \\ -1 & \text{若} n = 0 \end{cases} \tag{2.11}$$

   - CM_i ：特定月份中的第 i 個星期一的用電量。
   - CM_n ：表示該月份中其他星期一的用電量。

2. 處理離群值：使用了論文中的 Winsorization 方法來替換離群值。具體操作是使用 Least-Winsorized-Square 方法來將離群值替換為最接近的可接受值。

3. 資料正規化：如式 (2.12) 。

$$f(c_i) = \frac{c_i - \text{Min}(C)}{\text{Max}(C) - \text{Min}(C)} \tag{2.12}$$

   - c_i : 該用電量
   - $Min(C), Max(C)$ 表示用電量的最小值和最大值。

4. 處理資料不平衡問題：採用了 Random Under Bagging 的方法

   - 建立平衡的子集：首先，先把所有竊電的資料（少數類別）都放進每個子集裡。接著，從正常用戶（多數類別）的資料裡，隨機選取和竊電資料數量相同的樣本加入子集中。這樣每個子集裡，竊電和正常用戶的數量都是一樣的。
   - 重複這個過程：重複上述步驟，將原始的不平衡資料集轉換成多個平衡的子集。(在實際操作中，我們總共建立了 15 個子集，並取其平均來判斷其指標）

分析方法原理

CBLOF （Cluster-Based Local Outlier Factor）是一種用於異常檢測的無監督學習方法，特別適合處理不平衡資料和異常值的問題。該方法通過對資料進行聚類分析，識別出離群點，進而提升異常檢測的準確性。主要步驟如下：

1. 聚類分析：使用 K-means 或其他聚類算法，將資料集劃分成多個聚類。在實際操作中，我們選用 K-means 。
2. 離群因子計算：對於每個資料點，計算其與所屬聚類中心的距離，並結合聚類的大小，評估該資料點的離群因子，如式 (2.13) 。

$$\text{CBLOF}(p) = \begin{cases} |C_i| \cdot \min(\text{distance}(p, C_j)) & \text{where } p \in C_i, C_i \in SC \\ & \text{and } C_j \in LC \text{ for } j = 1 \text{ to } b \\ |C_i| \cdot \text{distance}(p, C_i) & \text{where } p \in C_i \text{ and } C_i \in LC \end{cases} \tag{2.13}$$

3. 對於屬於大聚類的資料點，其 CBLOF 值為該點到聚類中心的距離乘以聚類的大小。
4. 對於屬於小聚類的資料點，其 CBLOF 值為該點到最近的大聚類中心的距離乘以小聚類的大小。

CBLOF 值越高，表示該資料點越可能是異常值。我們可以根據這個值進行排序，並設定要抓出竊電者最異常的前 5 %(使用者可自訂)。由於電力公司的資源有限，無法逐一檢查所有用戶，因此這個方法能夠生成異常排名，讓電力公司的人員可以優先查詢最異常的用戶。

分析與驗證結果

CBLOF 模型驗證指標，見圖 2.11 所示。

- 一定的區分能力 (Average AUC: 0.67): 模型能夠比隨便猜測更好地區分竊電用戶和非竊電用戶。
- 高精確率 (Average Precision: 0.83): 模型預測的竊電用戶中，大多數確實在竊電。

- 低召回率 (Average Recall: 0.17): 模型僅檢測到部分竊電用戶，說明還有改進空間。這可能是因為在使用 CBLOF 時將 beta 設定得較大，讓模型可以更專注於極端異常值。此外，儘管有做不平衡處理，但資料不平衡的問題仍然可能影響召回率。
- 中等的 F1 分數 (Average F1 Score: 0.28)：表示模型在精確率和召回率之間的平衡還需要改進。
- 不錯的前 50 名預測精確率 (Average MAP@50: 0.98)：模型在前 50 個預測中表現很不錯，適合用於抓取最異常的竊電行為。而因為我們主要是想要看異常偷電的排名，所以 MAP@50 是我們看的主要指標。

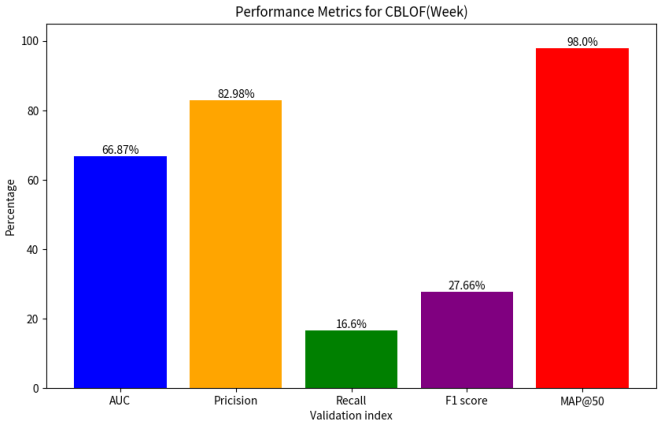


圖 2.11: CBLOF 模型指標

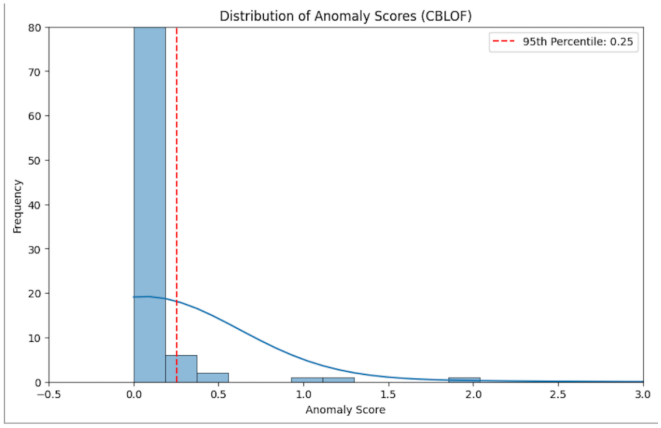


圖 2.12: CBLOF 模型預測結果與臨界值

3 輔助決策工具

本研究的輔助決策工具，建立一套系統來做偷電預測，系統設計稿見圖 3.1 所示，系統展示請見 (點此為系統連結)。

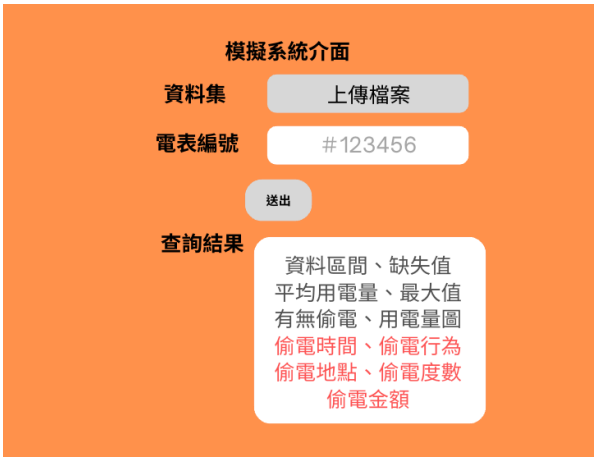


圖 3.1: 系統設計稿示意圖

3.1 使用說明書

本系統利用不同的統計方法來檢測和預測可疑的偷電行為。提供了結果展示和使用者查詢介面介面，分別針對整體用電資料分析及特定用戶分析。

3.1.1 使用注意事項

- 資料格式：確保匯入的電力資料格式正確，請使用圖 3.2 的資料格式，以避免影響分析結果的準確性。
- 方法選擇：不同的預測方法適用於不同的情境，建議根據實際需求選擇最合適的方法進行分析。

- 結果解讀：預測結果僅供參考，建議結合其他實際數據和情況進行綜合分析。

電表編號/日期	2014/1/1	2014/1/10	2014/1/11	2014/1/12	2014/1/13	2014/1/14
F52CFF361D1F87ACF5E1A9BD5D255A3C						
6F4919E1A9FEEF57C26BFF9DCEF97B1E	0.0	0.0	0.0	0.0	0.0	0.0
1FA9C27D9BE77C43A919891A6DCB40D0	0.0	0.0	0.0	0.0	0.0	0.0
065573157D11407F877066A70A25CFA8	5.5	0.0	0.0	0.0	6.02	3.6
3A8E03953795361C00A49C0D93434424	0.0	9.0	0.0	6.07	4.38	2.68

圖 3.2: 資料型態示意圖

3.1.2 結果展示介面：整體用電資料分析

1. 匯入資料。
2. 點擊「匯入資料」按鈕，選擇要上傳的電力資料文件，文件支援格式包括 CSV 和 Excel。
3. 資料上傳成功後，系統會自動在界面上顯示數據樣本，讓您檢查是否匯入正確。如圖 3.3 。

圖 3.3: 匯入資料介面

4. 從三種預測方法（Wide_CNN、Clof 和 Cblof）中，選擇一種進行分析，如圖，選擇預測方法後，點擊「Load」按鈕，系統將運行選定的模型進行分析。

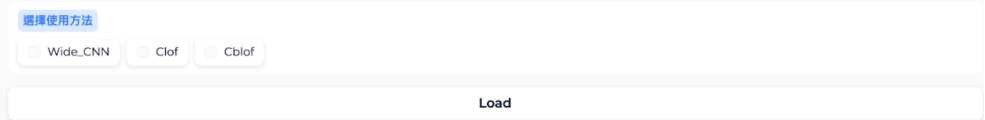


圖 3.4: 選擇預測方法

5. 分析完成後，系統會生成相關的指標值，如圖 2.8、2.10、2.12 (Wide_CNN $\geq$ 0.3, CLOF $\geq$ 1.28, CBLOF $\geq$ 0.4 。預測結果將依照

指標值進行排序，生成潛在的偷電用戶名單。點擊某一用戶可查看詳細的預測指標和用電行為分析報告。

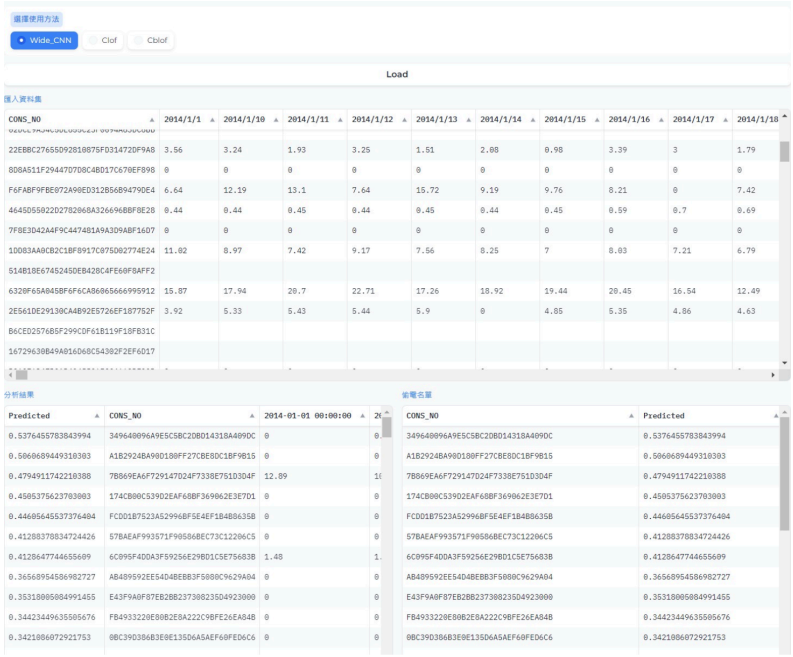


圖 3.5: 資料匯入與呈現

3.1.3 使用者查詢介面：特定用戶用電行為分析

1. 匯入資料：點擊「匯入資料」按鈕，選擇並上傳電力資料文件（支援 CSV 和 Excel 格式），如圖 3.6。

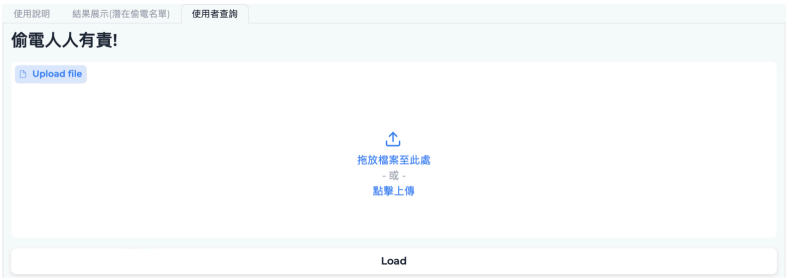


圖 3.6: 匯入資料介面

2. 選擇用戶與時間段，如圖，從顯示的資料中選擇您要分析的特定用戶。選擇需要分析的時間段，確保選擇的時間段內有足夠的數據供分析使用。

3. 選擇預測方法: 同樣提供三種預測方法（Wide_CNN、Clof 和 Cblof）作為選擇。

4. 選擇欲查詢的起始和結束日期，如圖 3.7。

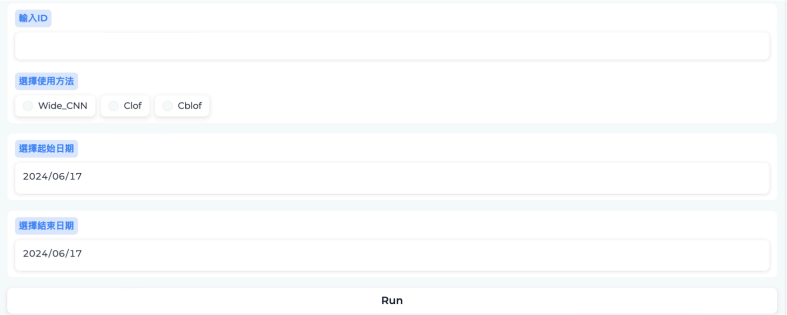


圖 3.7: 日期選定

5. 點擊「Run」按鈕，系統將針對選定的用戶和時間段進行分析，系統會顯示該段時間的用電行為分析結果，預測結果會指示該用戶是否存在偷電行為，並提供詳細的指標及相關資訊，如圖 3.8。

圖 3.8: 結果介面

客服支援

如在使用過程中遇到任何問題或需技術支援，請聯絡我們客服團隊，將竭誠為您提供幫助和支持。感謝您使用我們的電力偷竊預測應用程式。我們致力於為您提供高效且準確的預測工具，幫助您更好地管理和監控用電情況。希望您在使用過程中能夠獲得最佳的體驗。

4 結論

電力盜竊是一個全球性問題，每年造成的經濟損失達到數十億美元。這不僅對經濟造成巨大影響，還威脅到電力系統的穩定性和安全性。隨著竊電手段越來越多樣和隱秘，現有的偵測方法面臨巨大挑戰。在本研究中，我們建立了三種不同的模型來應對竊電問題，這些模型各自有其優缺點，並適用於不同的分析情境。這些模型的建立，旨在透過更精確和有效的方法來偵測竊電行為，從而減少因竊電帶來的經濟損失。除此，也開發了一個簡潔的系統，將這些模型應用於實際數據中，幫助電力公司更快速、更準確地識別潛在的竊電行為。此系統已經顯示出在實際應用中的巨大潛力，為電力公司提供了一個強有力的工具來打擊竊電。為了能增加本研究的實用性和使用上的便利性，我們下一代的產品將往以下三點的方向做改進，請客戶敬請期待。

1. 提高模型的準確性和穩定性：通過引入更多樣的數據和進一步優化模型參數，提升預測的準確性和穩定性。
2. 強化系統的實用性：改進系統的使用流程和界面，使其更加友好和高效，方便電力公司在實際操作中使用。
3. 綜合多種偵測方法：探索將更多不同的偵測方法綜合應用，以提高整體偵測效果，應對更加多樣化和隱秘的竊電手段。

總之，通過我們的研究和系統開發，為電力公司提供了一種有效的工具來應對日益嚴重的竊電問題，並為未來的進一步研究和改進打下了堅實的基礎。

附錄

組員分工名單

組別	姓名
Project Manager	劉家維、陳雅柔
Data Scientist	邱奕銓、楊家瑋、羅然莉
System Developer	張子恩、韓明澄

組員貢獻度

姓名	專案貢獻度
劉家維 (PM)	主要對外溝通窗口 (100%)、找文獻與方法 (80%)、專案進度管理 (40%)、上台報告與回答問題 (45%)、報告書撰寫(30%)、簡報製作(35%)、專案執行規劃(30%)、任務分配與協調 (10%)、文件管理 (5%)、主要對內溝通窗口(5%)
陳雅柔 (PM)	報告書統整 (100%)、文件管理 (95%)、主要對內溝通窗口(70%)、任務分配與協調 (60%)、簡報製作(55%)、專案執行規劃(50%)、專案進度管理 (40%)、上台報告與回答問題 (35%)、報告書撰寫(35%)、找文獻與方法 (20%)
邱奕銓 (DS)	資料日期格式整理 (100%)、寬深CNN離群值與正規化處理 (100%)、寬深CNN建模與指標新增 (60%)、CLOF離群值與正歸化處理 (30%)、CLOF缺失值日處理 (30%)、主要對內溝通窗口(20%)、報告書撰寫(20%)、

	任務分配與協調 (12.5%)、資料視覺化 (10%)、上台報告與回答問題 (7.5%)、簡報製作（6%）、CLOF建模與指標 (5%)、專案進度管理 (2.5%)、
楊家瑋 (DS)	寬深CNN缺失值日處理 (100%)、CLOF建模與指標 (85%)、CLOF離群值與正歸化處理 (70%)、CLOF缺失值日處理 (70%)、寬深CNN建模 (20%)、報告書撰寫（5%）、任務分配與協調 (2.5%)、專案進度管理 (2.5%)、簡報製作（1%）、
羅然莉 (DS)	週處理補缺值 (100%)、CBLOF的前處理、建模、驗證 (100%)、K-fold方法的建立 (100%)、資料觀察與視覺化 (90%)、寬深CNN建模 (20%)、專案執行規劃 (15%)、CLOF 研究 (10%)、專案進度管理 (10%)、上台報告與回答問題 (7.5%)、任務分配與協調 (5%)、報告書撰寫（5%）、簡報製作 (3%)
張子恩 (SL)	工具研究 gradio (70%)、系統介面設計 (70%)、導入模型 (70%)、建立選項 (60%)、輸出結果 (50%)、、顯示觀察結果 (50%)、輸入資料 (30%)、使用說明書撰寫 (20%)、專案執行規劃 (5%)、主要對內溝通窗口（5%）、上台報告與回答問題(4%)、報告書撰寫（2.5%、任務分配與協調 (2.5%)、專案進度管理 (2.5%)
韓明澄 (SL)	使用說明書撰寫 (80%)、輸入資料 (70%)、顯示觀察結果 (50%)、輸出結果 (50%)、建立選項 (40%)、導入模型 (30%)、工具研究 gradio (30%)、系統介面設計 (30%)、報告書撰寫（2.5%）、任務分配與協調(2.5%)、專案進度管理(2.5%)、上台報告與回答問題(1%)

會議紀錄

小組定期討論紀錄 (第一次報告準備期間、第二次報告準備期間、第三次報告準備期間)

開會日期	開會內容大綱
2024/03/25	1.方法研究 2.預期和最後目標 3.統整三次報告查核點 4.各組細分目標狀況與所需資源
2024/04/01	1.各組進度回報 ● PM:方法研究、回報文獻 ● DS:資料視覺化、遺失值處理、離群值處理 ● SL:系統工具研究 2.討論第一次開會報告內容
2024/04/08	1.各組進度回報 ● PM:第一次報告內容 ● DS:資料視覺化、遺失值處理、離群值處理、不平衡資料處理 ● SL:方法研究 2.實際演練第一次報告 3.討論系統介面雛形
2024/04/15	1.各組進度回報 ● PM:第一次報告問答討論 ● DS:方法研究 ● SL:提出網頁概念與雛形 2.檢討第一次報告 3.討論網頁概念與雛形
2024/04/22	1.各組進度回報 ● PM:方法研究、回報文獻 ● DS:建立 CNN、CLOF 模型 ● SL:展示系統初步介面 2.討論第二次報告內容
2024/04/29	1.各組進度回報 ● PM:無 ● DS:建立 CNN、CLOF 模型

	• SL：展示系統介面 2.討論第二次報告內容
2024/05/06	1.各組進度回報 • PM：第二次報告內容 • DS：缺失值重新處理、建立 CNN、CLOF 模型 • SL：系統工具研究 2.實際演練第二次報告內容
2024/05/13	1.各組進度回報 • PM：第二次報告問答討論 • DS：缺失值重新處理 • SL：討論系統介面日期輸入方式 2.檢討第二次報告
2024/05/20	1.各組進度回報 • PM：方法研究、回報文獻 • DS：優化 CNN 和 CLOF 模型、嘗試其他模型 • SL：系統介面UI 2.檢討第二次報告 3.討論第三次報告內容與報告書
2024/05/27	1.各組進度回報 • PM：無 • DS：週平均缺失值處理、問答討論 • SL：系統介面UI詳細資訊 2.討論第三次報告內容
2024/06/03	1.各組進度回報 • PM：第三次報告內容 • DS：建立 CBLOF 模型 • SL：系統介面完成 2.實際演練第三次報告

三次報告的 Q & A

第一次報告

問題	回答
偷電查獲比例是從哪裡拿到的資料	從立法院的報告書找出來的

甚麼是離群值	離群值是指過於偏離資料中心的資料，這邊我們使用的方式是把大於平均加 2 倍標準差的資料值統一降低到平均加 2 倍標準差。
離群值為什麼要調整	因為離群值因為數值過大，很容易去過度影響模型的判斷能力，因此需要做調整的動作，去降低離群值在模型判斷上的影響力，以免模型容易被雜訊影響。
有沒有可能調整提群值後反而使預測結果出現問題	有可能，因為調整後跟真實資料或多或少有些落差。但我們這邊根據實驗得出的結果是調整後的模型性能有明顯上升。
資料說明 資料那兩筆(有無偷電)怎麼選的	以是選取視覺化後差異較為明顯的兩筆資料，用以方便解釋週期性的現象。
為什麼有偷電，用電量高	推測原因可能是偷電者是因為有大量用電的需求，所以才會去做偷電這件事情。例如:挖取虛擬貨幣、或是在室內大麻種植等等。
甘特圖 如果 SL、DS 組成果不如預期怎麼辦？	PM 組會幫忙，且因會在開會時各組交流資訊，所以各組其實都有能力去幫忙其他組別的部份工作。

第二次報告

問題	回答
CNN 運作原理為何?	請參考 2.2.1「寬深CNN的分析方法原理」。
輸入的資料怎麼跑到深組件和寬組件？	在我們的前處理中，會將資料分成原始的一維資料和處理成類似圖片樣式的二維資料，並將他們放進模型做訓練。
MAP (mean Average Precision) 指標的意思?	MAP 指標為預測不同 label 的正確率的平均值。
離群值為什麼只取大於的沒有取小於的?	因為用電量恆正，因此最小值為 0。當資料突然變為 0 的情況下，就很有可能就是異常值，所以我們選擇不去對其做調整，

	因為這種值是我們判斷是否偷電的重要依據。

第三次報告

問題	回答
三個模型如何做評估優劣？	他們有一個共同使用的 F1-score，可透過這項指標來比較各個模型對整個資料集是否有足夠的能力去判別是否偷電
寬深 CNN 只是訓練時長比較久而已，怎麼會影響用戶的使用時間呢?	因為寬深 CNN 在資料的前處理上和另外兩個模型較為不同，需要花更多的時間將客戶放進來的資料集變成可以讓模型去做預測的形式
寬深 CNN 的運作原理是甚麼?	請參考 2.2.1「寬深 CNN 的分析方法原理」

與老師的討論紀錄

日期	問題	老師回應
3/8	會議記錄要記甚麼?	各組回報、事項可以分成已完成和未完成，以及誰完成和誰提出想法與回應。
3/8	個人貢獻有甚麼要點?	分工明確、要仔細確認每個人的意見，以及平常開會時就要做好紀錄，避免紛爭。
3/15	不一定能完成的目標該如何報告?	報告時可以和客戶溝通這個目標的可行性，以及告知目前的困難，順便跟

		他們要錢。
3/15	如何知道客戶的實際需求?	多和客戶進行溝通，報告進度，以免偏離他們的想像
3/15	如何知道計畫執行時可能會遇到甚麼困難?	通常要遇到才知道，但也可以多詢問各部們執行類似計畫時的經驗
3/15	專案執行計畫該如何制訂?	先大方向制定，再聚焦，並且滾動式調整
4/1	報告需要從頭開始說明嗎?	以面對客戶的角度，將他們想要和需要看到的內容報告出來即可
4/1	如果成員間發生矛盾該如何處理?	分別確認他們的狀況，尋求一個平衡點。
4/1	時程安排應該要怎樣才比較合理?	能如期完成就是好方法。

參考資料來源

Z. Zheng, Y. Yang, X. Niu, H. -N. Dai and Y. Zhou, "Wide and Deep Convolutional Neural Networks for Electricity-Theft Detection to Secure Smart Grids," in *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1606-1615, April 2018, doi: 10.1109/TII.2017.2785963.

H. M. Rouzbahani, H. Karimipour and L. Lei, "An Ensemble Deep Convolutional Neural Network Model for Electricity Theft Detection in Smart Grids," 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 2020, pp. 3637-3642, doi: 10.1109/SMC42975.2020.9282837.

Y. Peng et al., "Electricity Theft Detection in AMI Based on Clustering and Local Outlier Factor," in IEEE Access, vol. 9, pp. 107250-107259, 2021, doi: 10.1109/ACCESS.2021.3100980.