110-1 製造數據科學
Final Project

組別：第 17 組
指導教授：李家岩 教授
學生：
r09725006 劉心鈺
r09722016 劉庭安
r08725059 張煜柔
r10725014 莊芯瑜
Jan. 2022

Contents

1 Introduction

1.1 Background and Motivation

全球環保意識抬頭，許多企業響應聯合國所提出的 ESG(Environmental, Social and Governance) 經營指標，欲發展永續經營的理念，而興起對綠電的需求。近年，臺灣欲發展再生能源市場，推動了相關政策與法規，如設立 2025 年再生能源推廣目標由 1,000 萬瓩提高至 2,700 萬瓩以上的目標，並規定 300 家用電大戶企業須建立綠電典範[1]。

臺灣處於亞熱帶地區，整體年均日照接近 1,800 小時，尤其適合發展太陽能產業，惟地面型太陽能規劃可使用面積一直不甚理想，難以找到足夠的空地「種電」，不過在政府的推動下，現今民眾可以在屋頂架設太陽能板且售電與台電或經媒合的廠商，隨周邊生態逐漸成熟，家戶進入生產太陽能電的門檻有望逐步降低，可提供更多綠能到市場上，緩解供不應求的情形。

對臺灣企業而言，為擴大太陽能電使用量，企業可向民間收購或租賃住宅屋頂，建置太陽能電板發電。因此本研究欲從企業的需求角度出發，假設所有屋頂皆可架設太陽能板，希望了解指定區域最大可能供應電量。本研究首先以天氣觀測資料預測太陽能板容量因子[2]，再透過偵測範圍內屋頂面積，推算出指定區域的屋頂，其太陽光電最大可能年發電量。此研究結果可提供給企業在指定地點周圍可能獲得的太陽能光電量，以利企業規劃太陽能光電系統的建置策略。

1.2 Problem Definition

本研究假設有一公司欲收購某位置附近的房子屋頂以建置太陽能板，使企業每年皆可達到使用 P 度太陽能源的永續目標。本研究參考該位置過去一年的天氣狀況，幫助公司了解需設多大的裝置容量，以及收購多少面積、多大範圍的屋頂。

[1]國際再生能源發展趨勢與政策：https://www.re.org.tw/knowledge/more.aspx?cid=201&id=3966

[2]容量因子可理解為發電效率，為方便理解，本文將統稱容量因子為「發電效率」

2 Methodology

2.1 Research Framework

給定使用者預計建置太陽能板的目標位置，以及預計達成的年發電量，本系統結合太陽能日發電效率預測（capacity factor prediction）以及屋頂面機辨識（roof area recognition），計算達成目標年發電量所需收購的屋頂範圍，研究框架如圖 1。第 2.2 節為太陽能日發電效率預測，我們比較不同的預測模型，並保留補值的 KNN 模型，以及表現最好預測模型以供系統使用；第 2.3 節為屋頂面積辨識，我們透過 Google map API 獲取 road map 影像，並進一步辨識屋頂、計算屋頂面積；而第 2.3 節為系統設計，我們使用 Vue.js 作為前端框架，並使用 flask 做為後端框架，進行開發。

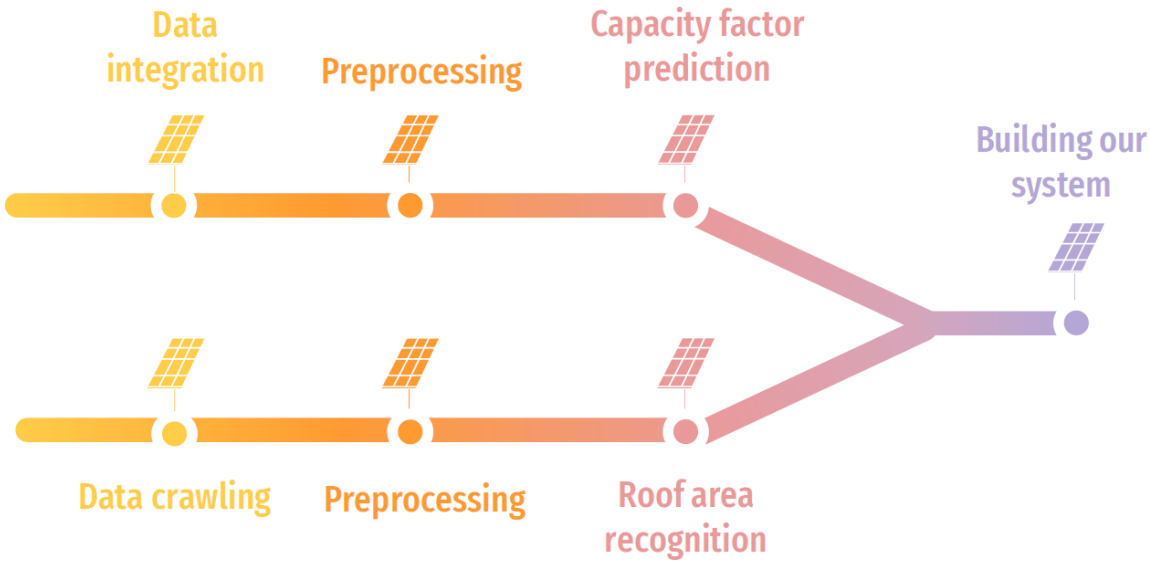

圖 1: 研究架構

2.2 太陽能日發電效率預測

2.2.1 資料前處理

在進行後續預測模型的訓練與驗證前，需檢視目前的資料型態、資料是否含有空值、變數之間的相關性以避免共線性等資料前處理。

- **類別變數處理**：首先檢視各變數的資料型態，若變數為類別變數須對其進行 get dummy，若有 n 個類別則需新增 n-1 個欄位表示其類別變數。若為數值變數則需確認讀入資料後、模型訓練前的變數型態為數值變數。

- **缺失值處理**：確認完資料型態後，檢視資料中的空值狀況。由於使用的天氣資料集包含自中央氣象局所屬有人站、自動氣象站搜集資料，可能會有大量空值存在，因部分天氣欄位僅能從中央氣象局所屬有人站搜集，如表 1 所示。因此，未解決空值情況，我們採用 KNN 演算法補值來處理缺失的天氣欄位資料。

氣象站	站號	觀測項目
中央氣象局地面氣象站	46XXXX	觀測時間、測站氣壓、海平面氣壓、 測站最高氣壓、測站最高氣壓時間、 測站最低氣壓、測站最低氣壓時間、 氣溫、最高氣溫、最高氣溫時間、 最低氣溫、最低氣溫時間、露點溫度、 相對溼度、最小相對溼度、 最小相對溼度時間、風速、風向、 最大陣風、最大陣風風向、 最大陣風風速時間、降水量、 降水時數、最大十分鐘降水量、 最大十分鐘降水量起始時間、 最大六十分鐘降水量、 最大六十分鐘降水量起始時間、 日照時數、日照率、全天空日射量、 能見度、A 型蒸發量、日最高紫外線指數、 日最高紫外線指數時間、總雲量
中央氣象局自動氣象站	C0XXXX	觀測時間、測站氣壓、測站最高氣壓、 測站最高氣壓時間、測站最低氣壓、 測站最低氣壓時間、氣溫、最高氣溫、 最高氣溫時間、最低氣溫、最低氣溫時間、 相對溼度、最小相對溼度、 最小相對溼度時間、風速、風向、 最大陣風、最大陣風風向、 最大陣風風速時間、降水量
中央氣象局自動雨量站	C1XXXX	降水量

表 1: 各類型氣象站提供的氣象參數

- **檢視是否有重複性高的自變數欄位**：為降低共線性問題，在建立模型前檢視自變數欄位間的相關係數，以避免同時選取相關性過高的欄位作為自變數。在此我們找出相關係數高於 0.9 或低於-0.9 的自變數組合，並從各組合中各選出一個變數放入自變數中。

2.2.2 太陽能發電效率模型訓練

本研究先根據天氣資料進行太陽能發電效率的預測。分析流程為：隨機切分訓練資料集與驗證資料集、使用訓練資料集做 cross validation 找出模型的最佳 hyperparameters 並對模型進行訓練、使用驗證資料集檢視模型預測結果（MAE、MSE）。預測模型主要可分為線性模型與樹狀模型。線性模型包含：OLS regression、LASSO regression, Ridge regression、Elastic Net。以下將分別解釋使用到的線性模型之方法。

- **OLS Regression**：可作為線性預測模型的 base line。目標為找到係數組合 β 最小化殘差平方總和，亦即

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 。$$

  在 2.2.1 節篩選掉相關性高的變數，以解決 OLS regression 的共線性問題。

- **LASSO regression**：為加入 L1 term 正則化（權重的絕對值和限制）後的線性迴歸變形，可避免 overfitting 的情況，亦即

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| 。$$

  由於 L1 term 的關係可行解區域為菱形，訓練過程中會使變數產生稀疏性，因此 LASSO regression 訓練出的模型相較於其他線性迴歸模型而言自變數數量較少。

- **Ridge regression**：為加入 L2 term 正則化（權重的平方和限制）後的線

性迴歸變形，同樣為避免 overfitting 的情況，亦即

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \circ$$

- **Elastic Net**：可以視為將 LASSO regression 與 Ridge regression（L1 regularization 與 L2 regularization）以線性組合的方式合起來：

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 \right) ,$$

  由此綜合 ridge 的優點（有效正規化優勢，預測效果較好）以及 LASSO 的優點（挑選變數，解決參數估計偏誤的問題）。

  考量到解不一定是合線性模型，也嘗試訓練樹狀的預測模型。樹狀模型具有良好的解釋性、強健性，並且不需線性模型一樣做正則化，也可以處理自變數間的交互作用關係等優點。本研究所使用到的樹狀預測模型包含：CART Regression tree 以及應用集成演算法的 Random Forest 與 Gradient Boosting。以下將分別解釋使用到的樹狀模型之方法。

- **CART Regression tree**：可作為樹狀分類模型的 base line，透過 CART tree 上的葉節點將訓練資料劃分成一個個相對簡單的群落，群落上再利用迴歸模型學習並預測值。

- **Random Forest**：是 Bagging 演算法，對訓練集樣本取後放回的重複抽樣（對欄與列都進行抽樣）以產生多個子資料集後，依序訓練成多個預測模型，再將所有模型的結果取平均，可助於降低模型的 variance。

- **Gradient Boosting**：是 Boosting 演算法，透過每次迭代的方式建立「弱模型」，每次迭代都基於前次的弱模型用梯度下降法去做優化、降低損失函數，以達到逐漸降低模型的 bias。

2.3 Google Road Map 屋頂面積辨識

由於本研究需計算特定區域內有多少的屋頂面積可作為太陽能發電區域，因此我們抓取 google 地圖並計算該區域內的屋頂面積。首先，本研究使用兩種方法讀取經緯度：讀取地址的經緯度、使用者自行輸入的經緯度，並以此經緯

度為中心點，向外擴展取得一區域範圍內的地圖，接著我們將地圖轉為灰階圖片，再使用線性濾波器銳利化圖像的邊界，再以影像二值化 Image Thresholding 的方法將圖片轉為僅有黑色與白色，然而白色面積會將河流或是高架橋等非屋頂的圖像也轉為白色，故我們額外計算這些圖像的面積，再將原白色面積扣除非屋頂圖像的面積，最後即可得區域內的屋頂面積。

舉寶藏巖國際藝術村為例，若以此地點為中心，抓取地圖後會包含河流面積，而一開始我們先轉為灰階圖片，如圖2，再轉為增強圖像的圖片，如圖 3，最後再計算白色的面積並扣除河流的面積，因此如圖 4 的白色部分為屋頂，而 5882.7 m^2 的面積則已扣除下面河流的面積區域。

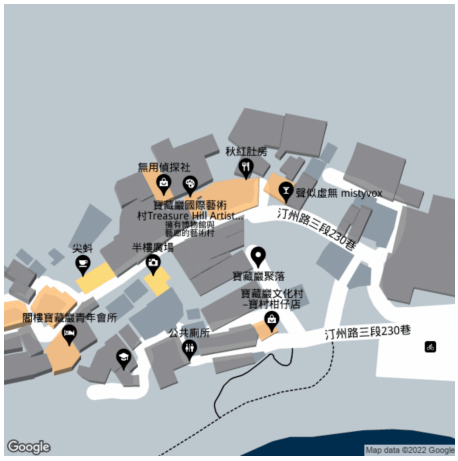

圖 2: Road Map 示例

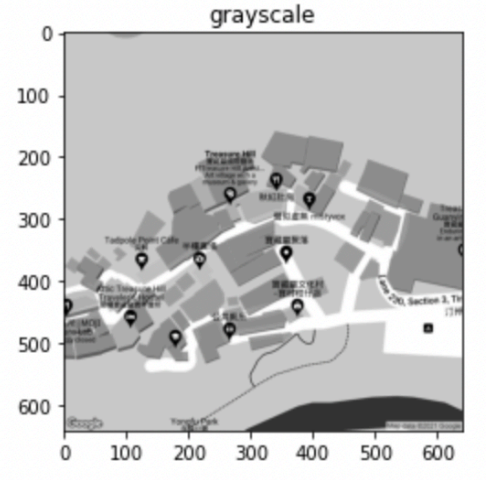

圖 3: 灰階圖片示例

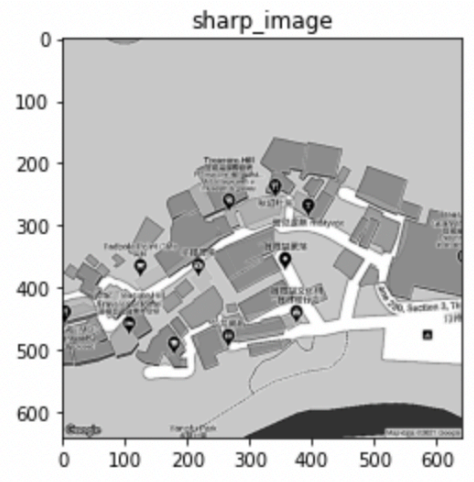

圖 4: 圖像增強示例

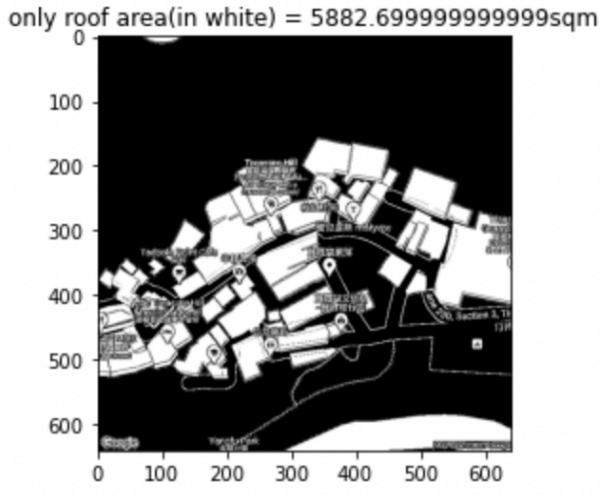

圖 5: 屋頂面積

2.4 系統設計

系統流程如圖 6。首先透過位置，我們可以計算離目標位置最近的天氣測站，並爬取此天氣測站過去一年（2020/10/01-2021/09/30）的日天氣資料，使用補值模型進行 KNN 補值，進一步透過預測模型得到每日的發電效率。由此我們可以進一步地計算要達到預期年發電量，所需建置多少太陽能板裝置容量，詳細公式如下，

$$K = \frac{P}{\sum_{t=1}^{365} \left(e_t \times 24\right)}$$

K 為所需的裝置容量（瓩）；e_t 為第 t 天的發電效率，$t = 1, ..., 365$；P 為預期達成的年發電量（度）。而 1 瓩的裝置容量約需 10 平方公尺的設置面積，因此由裝置容量我們可以進一步推得所需的屋頂面積。透過屋頂辨識模組，我們可以即時爬取目標位置附近的 google map 影像圖，因此可計算出需收購目標位置多大範圍的屋頂面積。

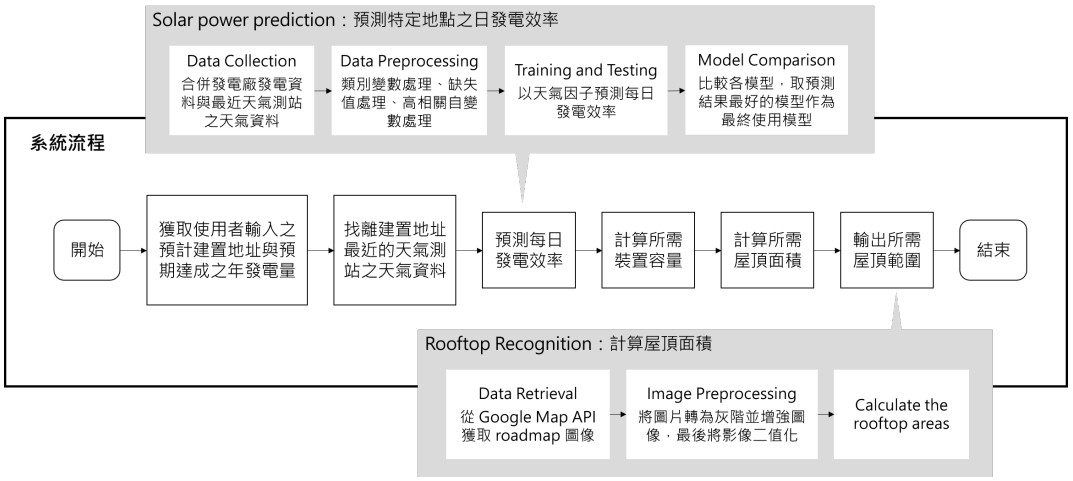

圖 6: 系統流程圖

3 Data Collection and Analysis Result

3.1 Data Collection

3.1.1 太陽能發電量

於政府資料開放平台上，取得太陽能發電量[3]與太陽能電廠裝置容量[4]資料。前者資料為每座太陽能發電廠每日的發電量，共計有 44 座太陽能發電廠資料，時間範圍為 2017/1/1 - 2021/9/30。而因為每座發電廠設置太陽能板數量不同造成發電量有明顯差異，我們透過後者發電廠裝置容量資料，將發電量轉換成發電效率，以此作為預測值。

3.1.2 天氣觀測資料

天氣資料取自觀測資料查詢平台[5]，會依照發電廠位置，取距離最近的自動觀測站月報表資料作為代表。月報表會記錄每日的天氣數值，如氣溫、累積降水量、日照時數等，共 36 個天氣特徵，以作為預測發電效率的特徵值。

3.1.3 Google Road Map

我們原本預計使用衛星影像圖作為屋頂辨識的資料，但是衛星影像圖的圖像含有過多資訊，在運算上會耗費過多時間，而因為我們希望結果能即時呈現在系統上，所以選擇使用 Google Road Map 的地圖種類，如圖 2。Google Road Map 為 Google 將衛星影像合理簡化的結果，只留下物體的輪廓，因此在辨識屋頂與計算屋頂面積時較容易且運算較快速。因此在本研究中，將使用 Google map API，獲取特定經緯度、特定範圍大小的 Road Map 影像圖，以進行屋頂辨識與屋頂面積計算。

3.2 Analysis

3.2.1 太陽能發電效率預測

本研究預測太陽能發電效率使用了線性模型與樹模型，針對樹模型的某些超參數調整，我們從隨機切分所得的訓練資料進行 10-fold cross validation，找

[3]台灣電力公司風力及太陽光電發電量資料：https://data.gov.tw/dataset/17140
[4]台灣電力公司太陽光電平均單位裝置容量統計表：https://data.gov.tw/dataset/29938
[5]自動測站觀測資料：https://e-service.cwb.gov.tw/HistoryDataQuery/index.jsp

出樹模型的最佳超參數，三個樹模型的驗證過程分別如圖7、圖8和圖9所示，我們挑選能最小化 negative mean squared error 的參數，因此最終 CART regression tree 採用 depth = 8、Random Forest 採用 max_depth = 10、GBDT 採用 1000 個 stages 數目。

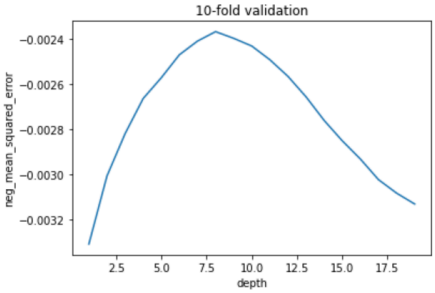

圖 7: regression tree 參數訓練

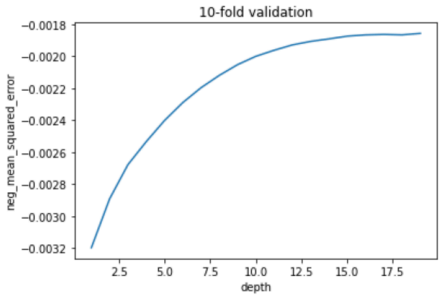

圖 8: random forest 參數訓練

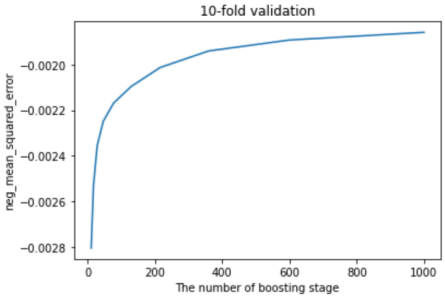

圖 9: GBDT 參數訓練

決定好超參數後，比較測試集於各個模型的表現狀況，並以 MSE 和 MAE 評價。模型測試結果以 MSE 小至大排列如表2和圖10，可以發現三個樹模型皆表現優於線性模型，其中以 Random Forest 表現最佳。而在線性模型中，有加入平方項的 OLS 表現最佳，此結果顯示天氣與太陽能板發電效率的關係僅以線性關係可能難以描述。

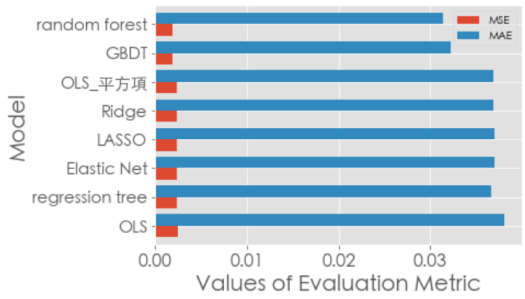

圖 10: 模型測試結果的 MSE 和 MAE

表 2: 模型測試結果的 MSE 和 MAE

model	MSE	MAE
Random Forest	0.0314	0.0018
GBDT	0.0323	0.0018
regression tree	0.0367	0.0024
OLS_squared	0.0369	0.0024
Ridge regression	0.0369	0.0024
LASSO	0.037	0.0024
Elastic Net	0.037	0.0024
OLS	0.0381	0.0025

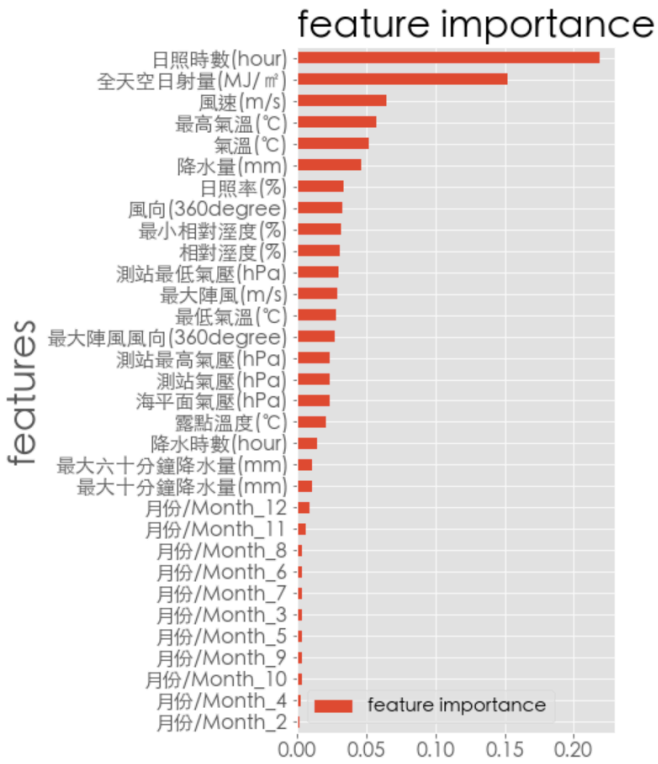

圖 11: random forest 的重要變數排序

接著從表現最佳的 Random Forest 討論變數重要性，此處以 Gini importance 的作為衡量方法，即能降低越多 impurity 的變數越重要。將重要性大至小排列如圖11，發現日照時數、全天空日射量是模型篩選最為重要的變數，此二天氣變數包含了陽光的照射時間長短、照射角度等資訊，顯示陽光是主要影響太陽能板發現效率的因子。

3.2.2 Google Road Map 屋頂辨識

Google Road Map 屋頂辨識的驗證，我們比較了程式算出的結果，以及從 Google Map 上手動框取建物的結果。圖 12 為單一建物驗證結果，程式計算為 4315.88 平方公尺，而從 Google Map 手動框取建物結果為 4303.71 平方公尺，兩者差異為 12.71 平方公里，差距並不大。圖 13 為多建物驗證結果，程式辨識出在此區域屋頂共有 5863.65 平方公尺，而人工量測 Google Map 上的屋頂面積，總面積為 5661.74 平方公尺，兩者差距不大。

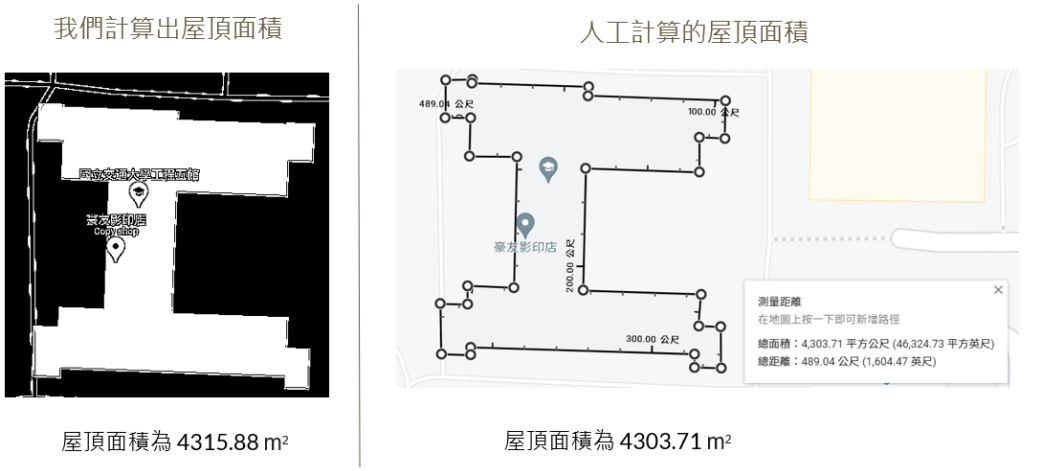

圖 12: 屋頂面積驗證-單一建物

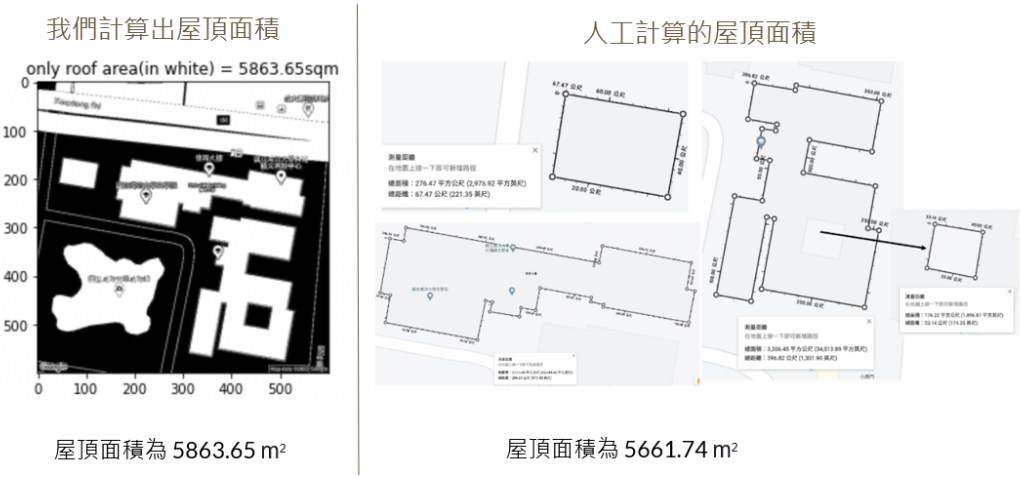

圖 13: 屋頂面積驗證-多建物

3.2.3 系統設計

系統介面如圖 14，左邊為參數輸入，右側為計算結果。使用者需在左邊區塊，輸入預計設置發電板的目標位置，以及預計達成的年發電量（度），其中位置可輸入經緯度或地址，或是透過滑動地圖、移動藍色圖標設定。按下左下的藍色按鈕「屋頂面積計算」，可得結果如圖 15。以此範例來說，經緯度設在於新竹，經緯度為 $(24.803647,\ 120.966585)$，預計達成 $2022,102$ 度年發電量，則需收購約 $13,878.97$ 平方公尺面積的屋頂，建置 $1,387.897$ 瓩的裝置容量，圖片顯示的以 $(24.803647,\ 120.966585)$ 當原點（圖片左上點），所需收購的屋頂範圍。

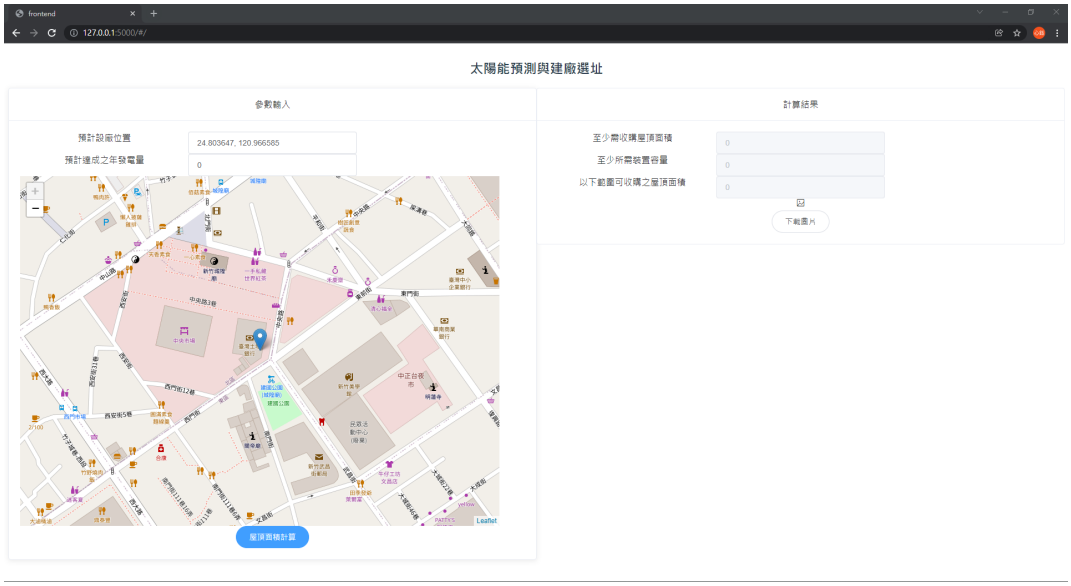

圖 14: 系統介面

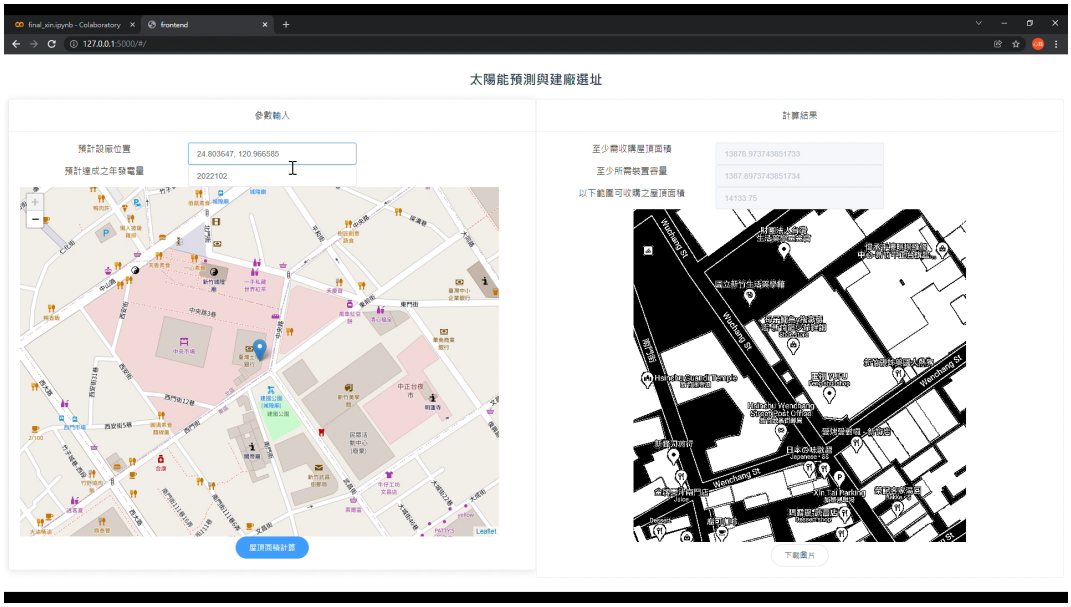

圖 15: 系統結果-新竹

3.3 Interpretation

透過精準的太陽能日發電效率預測與屋頂辨識，使用者可以任意輸入預計設置發電板位置與太陽能目標年發電量，系統會計算出所需收購的屋頂面積與建議收購範圍。以圖 15 舉例，若某一公司希望能收購新竹市中心附近 (24.803647, 120.966585) 的住宅屋頂，以建置太陽能發電系統，提供公司綠能使用，並且希望能達到 2022, 102 度的年發電量，則根據過去一年新竹市中心的天氣資料，我們建議至少需收購 13, 878.97 平方公尺的屋頂面積，才能達到 2022, 102 度年發電量，而推薦收購屋頂範圍如白色區域。

此結果會依據公司所選擇的預計設置位置之天氣、房屋大小、密度而有所不同，可以幫助公司比較不同位置所需收購範圍與難易度。舉例來說，如果公司亦考慮與台南成大合作，建設太陽能板於各系館屋頂，我們的系統可以幫助公司比較兩者的差異。預計達成的年發電量，結果如圖 16，建議至少需收購 11, 945.21 平方公尺的屋頂面積，所需收購範圍如畫面所示（截圖限制只截取到部分）。

圖 16: 系統結果-台南

由此可以發現，台南發電效率較高，因此需收購的屋頂面積比新竹來說較小，但因為所選地點房屋較分散，所以整體收購範圍較大。公司可以根據自己的考量，比較兩者差距，選擇建置太陽能板位置。

4 Conclusion

本研究結合太陽能日發電效率預測與屋頂面積辨識，支援企業進行再生能源建置決策。在太陽能日發電效率預測的部分，採用 Tree-based models 的預測結果較 linear models 好，其中以 Random Forest 預測效果最佳，此模型將於後續系統使用；而屋頂面積辨識的部分，使用 Google Road Map 辨識與計算，求得屋頂面積近似值；最後結合兩者的分析，建置成系統，方便提供企業使用，以協助公司做屋頂購買決策。