機器學習與大數據分析技術

Supervised Learning for Classification

羅勻瑄、黃世君、劉定睿

日期：2026-04-29

目錄

Download a dataset from Kaggle

Phishing website Detector：
https://www.kaggle.com/datasets/eswarchandt/phishing-website-detector
(https://www.kaggle.com/datasets/eswarchandt/phishing-website-detector)

本資料集共包含 11,055 筆樣本與 30 個特徵欄位，目標欄位為 Result，標籤分布為釣魚網站（-1）約 4,898 筆、正常網站（1）約 6,157 筆，類別比例約為 44:56，分布尚屬平衡，無需進行特殊的類別不平衡處理。

項目	內容
總筆數	11,055
特徵數量	30
釣魚網站（-1）	4,898 筆
正常網站（1）	6,157 筆
類別比例	約 44:56

Perform data preprocessing (cleaning, normalization if needed)

在模型訓練前，我們執行了以下步驟以確保數據品質：

1. 缺失值檢查： 經檢查該數據集完整，無缺失值。
2. 標籤轉換： 將原始標籤 -1 與 1 轉換為 0（釣魚）與 1（正常），以便計算 ROC/AUC 指標。
3. 數據分割： 將數據集依 80/20 比例分割為訓練集（Training Set）與測試集（Test Set）。
4. 特徵標準化： 使用 StandardScaler 進行縮放，這對於確保邏輯回歸等演算法的收斂至關重要。

本資料集的特徵雖皆為離散數值（-1、0、1），但為確保演算法的收斂穩定性，仍進行了 StandardScaler 標準化。其中，Logistic Regression 使用梯度下降法進行參數優化，若各特徵的數值尺度不一致，會導致梯度更新不穩定、收斂速度變慢，因此標準化

對其至關重要。
Random Forest 基於決策樹的分裂機制，本身對特徵尺度不敏感，標準化對其效能影響有限，但統一處理可確保兩個模型使用相同的輸入格式，便於公平比較。

Apply classification algorithms learned in class

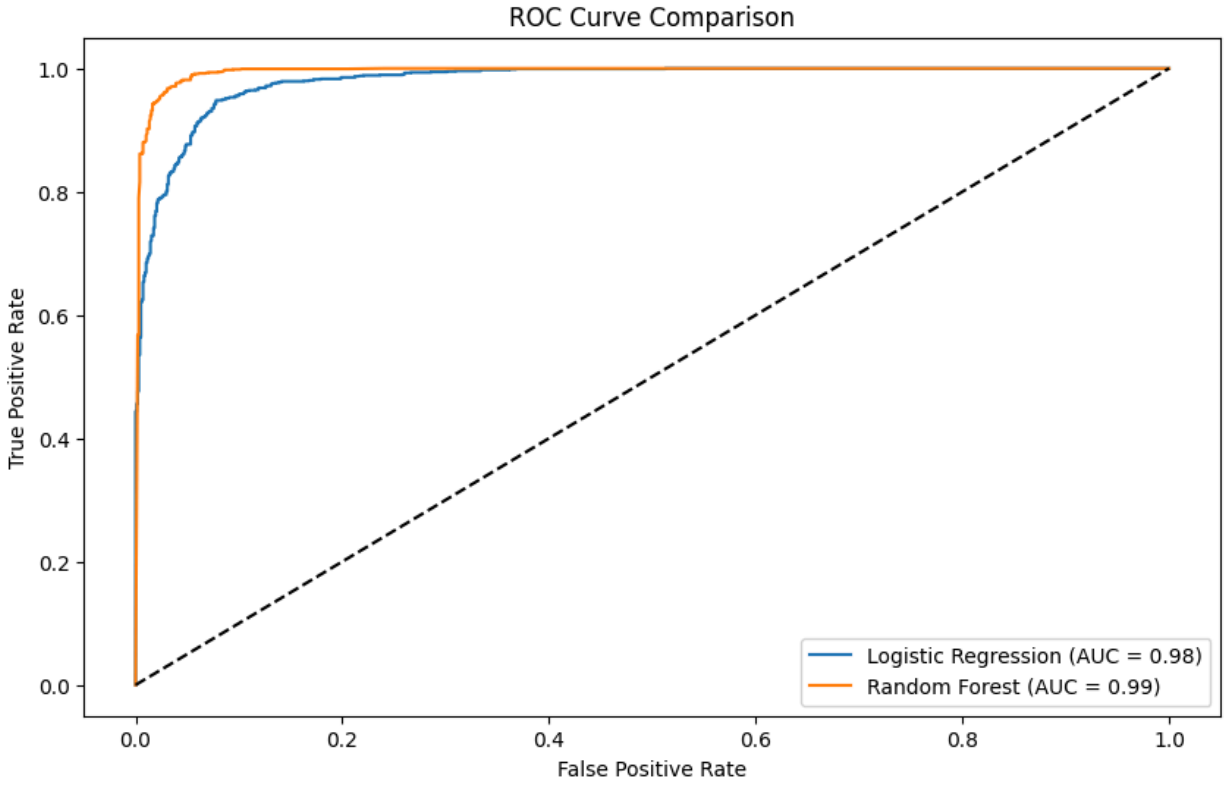
- Logistic regression, decision tree, random forest, SVM, KNN
- Choose two algorithms and compare their results.

我們選擇邏輯回歸 (Logistic Regression) 與 隨機森林 (Random Forest) 進行對比實驗，選擇原因如下：

Logistic Regression（邏輯回歸）
邏輯回歸是最基礎的線性分類模型，透過 Sigmoid 函數將線性組合的特徵值映射至 [0, 1] 的機率範圍，以 0.5 為決策邊界進行二元分類。其優點為訓練速度快、可解釋性高，輸出的係數可直接反映各特徵對預測結果的影響方向與強度，適合作為基準模型（Baseline）與其他複雜模型進行比較。

Random Forest（隨機森林）
隨機森林是一種集成學習（Ensemble Learning）方法，屬於 Bagging 技術的代表。其核心概念是透過對訓練資料進行有放回的隨機抽樣（Bootstrap Sampling），建立多棵獨立的決策樹，最終以多數決的方式整合各棵樹的預測結果。隨機森林能夠有效捕捉特徵之間的非線性關係，且內建特徵重要性（Feature Importance）分析功能，對 Data Insights 的探討極具參考價值。

一者為線性模型，一者為非線性集成模型，兩者在模型複雜度、可解釋性與預測能力上各有取捨，能夠充分展示不同演算法設計哲學對分類任務的影響。



上圖為兩個模型的 ROC 曲線比較。ROC 曲線以「False Positive Rate（誤報率）」為橫軸、「True Positive Rate（召回率）」為縱軸，曲線越靠近左上角代表模型分類能力越強。虛線對角線代表隨機猜測的基準線（AUC = 0.5）。

由圖可知，兩個模型的曲線均大幅高於基準線，表現優異。其中 Random Forest（AUC = 0.99）的曲線更緊貼左上角，優於 Logistic Regression（AUC = 0.98），代表在各種不同的決策閾值下，Random Forest 均能維持更高的分類準確性。

Model Evaluation

Random Forest Classification Report

Class	Precision	Recall	F1-score	Support
0	0.97	0.96	0.96	976
1	0.97	0.98	0.97	1235
Accuracy			0.97	2211
Macro Avg	0.97	0.97	0.97	2211
Weighted Avg	0.97	0.97	0.97	2211

Logistic Regression Classification Report

Class	Precision	Recall	F1-score	Support
0	0.94	0.91	0.92	976
1	0.93	0.95	0.94	1235
Accuracy			0.93	2211
Macro Avg	0.93	0.93	0.93	2211
Weighted Avg	0.93	0.93	0.93	2211

以上兩份 Classification Report 分別呈現了兩個模型針對各類別的詳細評估結果，其中 Class 0 代表釣魚網站、Class 1 代表正常網站。

由 Random Forest 的報告可知，兩類別的 Precision 與 Recall 均達到 0.96 以上，顯示模型對釣魚與正常網站的判斷皆十分準確。

Logistic Regression 的整體表現稍低，但對 Class 1（正常網站）的 Recall 達 0.95，代表絕大多數正常網站能被正確識別，誤判為釣魚網站的比例極低。

本研究使用 **Accuracy**、**Precision**、**Recall**、**F1-score**、**AUC**、**ROC-Curve** 作為模型評估指標。

	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.933514	0.929700	0.953036	0.941224	0.980449
Random Forest	0.969245	0.968675	0.976518	0.972581	0.994948

Cross-Validation

- Apply K-Fold cross-validation (e.g., K=5)
- Examine mean and standard deviation of each fold

Random Forest - 5-Fold CV Accuracy 結果：

- 各折分數: [0.96924469 0.97014925 0.97060154 0.96969697 0.97375566]
- 平均值 (Mean): 0.9707
- 標準差 (Std Deviation): 0.0016

由結果可知，極低的標準差顯示模型在不同的數據分割下表現穩定，證明了模型並非僅是對特定數據過擬合，而是具備強健的泛化能力。

Logistic Regression - 5-Fold CV Accuracy 結果：

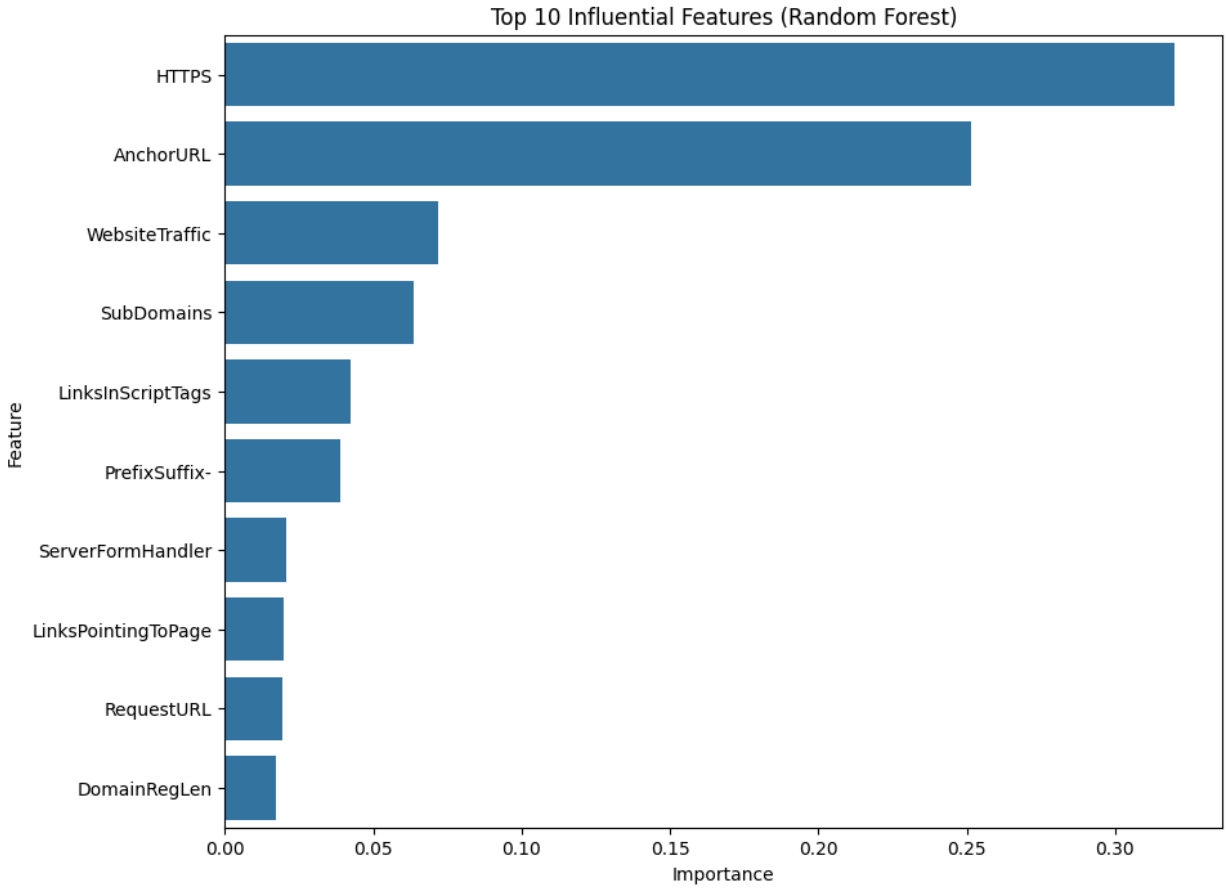
- 各折分數: [0.93351425 0.92175486 0.93080054 0.92175486 0.92986425]
- 平均值 (Mean): 0.9275
- 標準差 (Std Deviation): 0.0049

Logistic Regression 的標準差（0.0049）略高於 Random Forest（0.0016），顯示其在不同資料分割下的穩定性稍低，但整體仍屬穩定範圍。兩模型的 CV 標準差皆小於 0.01，共同說明兩者均未出現嚴重的過擬合現象，具備良好的泛化能力。

Data Insights

1. Are there any features that appear to strongly influence the outcome?

透過隨機森林的 `feature_importances_` 屬性，我們分析了各特徵對預測結果的貢獻程度，並繪製出前 10 大關鍵特徵的重要性圖表。



結果顯示，HTTPS（即網站是否使用加密連線）為影響力最高的特徵，重要性分數遠高於其他特徵，佔據主導地位。其次是 AnchorURL（網頁中超連結錨點的行為特徵），兩者合計貢獻了模型判斷的核心依據。其餘影響力較高的特徵依序包括 `WebsiteTraffic`、`SubDomains`、`LinksInScriptTags` 等。

這樣的結果與資安領域的實務知識高度吻合：

- HTTPS 是釣魚網站最常被忽略的技術特徵。
  合法網站通常會申請 SSL 憑證並啟用 HTTPS，而釣魚網站由於架設倉促、成本考量，往往使用未加密的 HTTP 連線，因此 HTTPS 狀態成為最具判別力的指標。
- AnchorURL 反映了頁面中超連結的指向行為。
  釣魚網站常在頁面內嵌大量指向外部惡意網域的連結，或使用與顯示文字不符的隱藏連結，此行為模式極具識別性。

HTTPS 與 AnchorURL 是本資料集中判斷釣魚網站最關鍵的兩個特徵，建議在實際的網路安全偵測系統中優先採用這兩項指標。

2. Compare the performance of different models

我們針對 Logistic Regression 與 Random Forest 兩個模型，從多個評估維度進行全面比較：

指標	Logistic Regression	Random Forest
Accuracy	0.9335	0.9692
Precision	0.9297	0.9687
Recall	0.9530	0.9765
F1-score	0.9412	0.9726
AUC	0.9804	0.9949

分析如下：

準確率（Accuracy）方面，Random Forest 以 96.92% 領先 Logistic Regression 的 93.35%，差距約 3.6 個百分點。

Precision 與 Recall 方面，Random Forest 在兩項指標上均優於 Logistic Regression。值得注意的是，在釣魚網站偵測的實務場景中，Recall（召回率）尤為重要——Recall 代表「所有真正的釣魚網站中，有多少被成功偵測出來」。若 Recall 過低，代表有大量釣魚網站漏網，對使用者造成實際危害。Random Forest 的 Recall（0.9765）高於 Logistic Regression（0.9530），在安全性考量上更具優勢。

AUC 方面，Random Forest（0.9949）幾乎接近完美分類器（AUC = 1.0），顯示其在不同決策閾值下均能維持極高的分類能力；Logistic Regression（0.9804）雖然稍低，但也屬於優秀等級。

總結來說，Random Forest 在所有指標上均優於 Logistic Regression，原因在於 Random Forest 是集成學習方法，透過組合多棵決策樹來降低變異度，能夠捕捉特徵之間的非線性關係；而 Logistic Regression 假設特徵與目標之間存在線性關係，對於本資料集中複雜的特徵交互作用，表達能力相對有限。

若以偵測準確性為優先考量，建議選用 Random Forest；若需要模型具備較高可解釋性（例如向非技術人員說明判斷依據），Logistic Regression 仍是可接受的選擇。

3. Discuss whether the model suffers from overfitting

我們透過比較兩個模型在訓練集與測試集上的準確率，以及 5-Fold Cross-Validation 的標準差，來評估是否存在過度擬合問題。

Random Forest：

	數值
訓練集準確率	0.9911
測試集準確率	0.9692
差異	0.0218
CV Mean	0.9707
CV Std	0.0016

訓練集與測試集準確率差距僅約 2.2%，屬於合理範圍。Random Forest 作為一種 Bagging 技術，透過對訓練資料進行有放回抽樣（Bootstrap Sampling），並平行組合多棵決策樹的預測結果，有效降低了模型的變異度（Variance），從而抑制了過度擬合問題。CV 標準差極低（0.0016），進一步確認模型在不同資料切割下表現高度一致，泛化能力強健。

Logistic Regression：

	數值
訓練集準確率	0.9267
測試集準確率	0.9335
差異	0.0068
CV Mean	0.9275
CV Std	0.0049

Logistic Regression 本身具有較低的模型複雜度，其假設空間受限於線性決策邊界，因此天生較不容易過擬合。從 Cross-Validation 結果來看，CV Mean 為 0.9275、Std 為 0.0049，標準差雖略高於 Random Forest，但仍在合理範圍內，顯示模型穩定性良好，同樣未出現明顯的過度擬合現象。

值得注意的是，Logistic Regression 的訓練集準確率（0.9267）略低於測試集（0.9335），差異為 0.0068。
此現象在機器學習中並不常見，但有幾個解釋：

1. **資料分割的隨機性：** 本實驗以 `random_state=42` 進行 80/20 分割，測試集的樣本分布恰好對 LR 的線性決策邊界較為友善，導致測試集表現略優於訓練集。

2. **StandardScaler 的 fit 範圍：** Scaler 僅對訓練集 fit，再 transform 測試集。若測試集的特徵分布恰好更接近標準常態分布，LR 的梯度下降收斂效果可能在測試集上表現更佳。

3. **模型本身的低複雜度：** LR 為線性模型，本身極不容易過擬合，訓練集與測試集的準確率差距極小（<1%），整體仍屬正常範圍。

改善方向： 若要使結果更穩定，可使用 `StratifiedKFold` 取代單次 train_test_split，確保每次分割的類別比例一致，減少隨機性的影響。

綜合比較：

兩個模型均未出現嚴重的過度擬合，其中 Random Forest 的 CV 標準差（0.0016）優於 Logistic Regression（0.0049），顯示 Random Forest 在不同資料分割下更為穩定。整體而言，兩個模型皆具備良好的泛化能力，能夠有效應用於真實世界的釣魚網站偵測任務。

總結來說，本次實驗的兩個模型均無過度擬合疑慮，Random Forest 的表現更為穩定，是本任務的最佳選擇。