機器學習期末報告

主題: PhiUSIIL 釣魚網站 URL 的準確預測與開發

組員:葉宇倫、楊書暐

摘要

本研究使用 UCI machine learning repository 上的 PhiUSIIL Phishing URL (Website)資料集。第一部分是資料的探索式分析，第二部分是使用不同變數與整個資料集的建模預測，以用來與最後一個部分的演算法建模預測做比較。

第一章 緒論

第一節 研究動機與目的

隨著新冠疫情的爆發與全球大流行，各國政府、企業與個人紛紛加速採用數位轉型來應對社會運作與經濟活動的中斷。這一過程促使線上交易、遠距工作、虛擬會議與電子商務等需求激增，網路空間的重要性與依賴性也隨之大幅提高。然而，隨著網路活動的增加，釣魚網址的攻擊頻率亦呈現驚人的上升趨勢，對於個人隱私、財務安全與企業機密構成了前所未有的威脅。尤其是在疫情期間，犯罪者利用大眾對疫情資訊、政府補助或疫苗接種等緊急需求的渴望，進一步提高了釣魚網站的成功率，甚至針對弱勢族群和初次使用數位工具的用戶進行攻擊，造成了廣泛的社會與經濟損失。因此，為有效防範此類網路威脅，本研究旨在深入探討並尋找釣魚網站篩選的有效特徵以及探討各類特徵對釣魚網站的影響，以提升偵測準確率，從而降低網路安全風險。本研究的成果將為未來網路安全防護措施的制定提供重要參考，進一步保障數位時代中用戶的安全與信任。

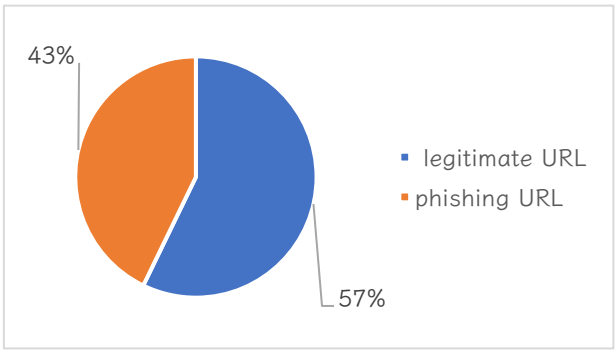
第二節 研究流程

我們先進行探索式資料分析找出可能是重要的特徵去建模，再建立非人造變數資料集與人造變數資料集(原資料集)各兩個模型，共三個模型去做預測比較，最後與論文內演算法建模做比較，並總結一些成果，提出進一步的建議。

第二章 探索式資料分析

第一節 變數介紹與總結

本資料集有 235798 筆觀察值，56 個特徵，以及目標二元變數 label，0 表示釣魚 URL，1 代表非釣魚 URL。該資料集無缺失值，且 label 大致上是均衡的，見下方圖一。



圖一、label 分布圓餅圖

變數名稱	說明	變數名稱	說明	變數名稱	說明	變數名稱	說明
HasObfuscation	是否存在混淆	HasPasswordField	是否存在密碼欄位	NoOfObfuscatedChar	混淆字元的數量	label	是否為釣魚網站
HasFavicon	是否有網站圖標（Favicon）	HasCopyrightInfo	是否包含版權資訊	NoOfAmpersandInURL	URL 中的「&」符號數量	NoOfOtherSpecialCharsInURL	URL 中其他特殊字元的數量
HasDescription	是否包含描述	HasTitle	是否有標題	NoOfLettersInURL	URL 中字母的數量	NoOfURLRedirect	URL 的重導次數
HasExternalFormSubmit	是否使用外部表單提交	IsDomainIP	是否使用 IP 當作域名	NoOfDegitsInURL	URL 中數字的數量	NoOfSelfRedirect	自我重導的次數
HasSocialNet	是否與社交網絡相關	IsHTTPS	是否使用 HTTPS	NoOfEqualsInURL	URL 中等號（=）的數量	NoOfEmptyRef	空引用的數量
HasSubmitButton	是否有提交按鈕	IsResponsive	是否響應式	NoOfSelfRef	自引用的次數	NoOfPopup	彈出視窗的數量
HasHiddenFields	是否存在隱藏欄位	NoOfSubDomain	子域名的數量	NoOfQMarkInURL	URL 中問號（?）的數量	NoOfiFrame	iFrame 的數量
NoOfImage	圖片數量	TLD	頂級域名（Top Level Domain）	URLSimilarityIndex	URL 相似性指標	DomainTitleMatchScore	域名與標題的匹配分數
NoOfCSS	CSS 檔案的數量	Title	標題	CharContinuationRate	字符延續率	URLTitleMatchScore	URL 與標題的匹配分數
NoOfJS	JavaScript 檔案的數量	URLLength	URL 的長度	TLDLegitimateProb	頂級域名的合法概率	ObfuscationRatio	混淆比例
NoOfExternalRef	外部引用的數量	DomainLength	域名的長度	URLCharProb	URL 字符的概率	Bank	與銀行相關的標記
FILENAME	文件名稱	TLDLength	頂級域名的長度	LetterRatioInURL	URL 中字母比例	Pay	是否與支付相關
URL	URL 本身	LineOfCode	代碼行數	DegitRatioInURL	URL 中數字比例	Crypto	與加密相關的標記
Domain	域名	LargestLineLength	最長代碼行的長度	SpacialCharRatioInURL	URL 中特殊字符比例	Robots	是否與機器人文件相關

類別變數	
變數名稱	唯一值
FILENAME	235795
URL	235370
Domain	220086
TLD	695
Title	197874

數值變數 及 二元變數				
變數名稱	平均	標準差	最小值	最大值
URLLength	34.5731	41.3142	13	6097
DomainLength	21.4704	9.1508	4	110
IsDomainIP	0.0027	0.0519	0	1
URLSimilarityIndex	78.4308	28.9761	0.155574	100
CharContinuationRate	0.8455	0.2166	0	1
TLDLegitimateProb	0.2604	0.2516	0	0.522907
URLCharProb	0.0557	0.0106	0.001083	0.090824
TLDLength	2.7645	0.5997	2	13
NoOfSubDomain	1.1648	0.601	0	10
HasObfuscation	0.0021	0.0453	0	1
NoOfObfuscatedChar	0.0249	1.8762	0	447
ObfuscationRatio	0.0001	0.0038	0	0.348
NoOfLettersInURL	19.4289	29.0903	0	5191
LetterRatioInURL	0.5159	0.1233	0	0.926
NoOfDegitsInURL	1.881	11.8867	0	2011
DegitRatioInURL	0.0286	0.0709	0	0.684
NoOfEqualsInURL	0.0622	0.9347	0	176
NoOfQMarkInURL	0.0294	0.1935	0	4
NoOfAmpersandInURL	0.0251	0.8364	0	149
NoOfOtherSpecialCharsInURL	2.3402	3.5276	0	499
SpacialCharRatioInURL	0.0633	0.0324	0	0.397
IsHTTPS	0.7826	0.4125	0	1
LineOfCode	1141.9	3419.951	2	442666
LargestLineLength	12789.53	152201.1	22	13975732

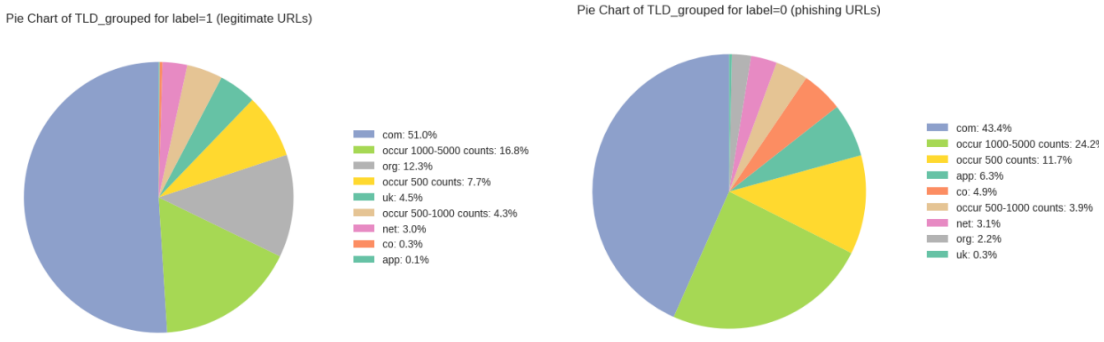
HasTitle	0.8613	0.3457	0	1
DomainTitleMatchScore	50.1314	49.677	0	100
URLTitleMatchScore	52.1221	49.6006	0	100
HasFavicon	0.3618	0.4805	0	1
Robots	0.2665	0.4422	0	1
IsResponsive	0.6245	0.4842	0	1
NoOfURLRedirect	0.1334	0.34	0	1
NoOfSelfRedirect	0.0401	0.1962	0	1
HasDescription	0.4402	0.4964	0	1
NoOfPopup	0.2218	3.8705	0	602
NoOfiFrame	1.5886	5.7626	0	1602
HasExternalFormSubmit	0.044	0.2051	0	1
HasSocialNet	0.4566	0.4981	0	1
HasSubmitButton	0.4143	0.4926	0	1
HasHiddenFields	0.3778	0.4848	0	1
HasPasswordField	0.1023	0.303	0	1
Bank	0.1271	0.3331	0	1
Pay	0.237	0.4252	0	1
Crypto	0.0235	0.1514	0	1
HasCopyrightInfo	0.4868	0.4998	0	1
NoOfImage	26.0757	79.4118	0	8956
NoOfCSS	6.3331	74.8663	0	35820
NoOfJS	10.5223	22.3122	0	6957
NoOfSelfRef	65.0711	176.6875	0	27397
NoOfEmptyRef	2.3776	17.6411	0	4887
NoOfExternalRef	49.2625	161.0274	0	27516
label	0.5719	0.4948	0	1

表一、變數介紹表格

第二節 數據圖片與特徵發現

因為其他類別變數個數太多，因此只留下不同類別數僅有 695 的 TLD，再根據他們的 count 進行分類編碼，並使用卡方檢定得到這些類別非獨立，建模時會使用編碼過的 TLD 做成新特徵取代原本 TLD 變數。

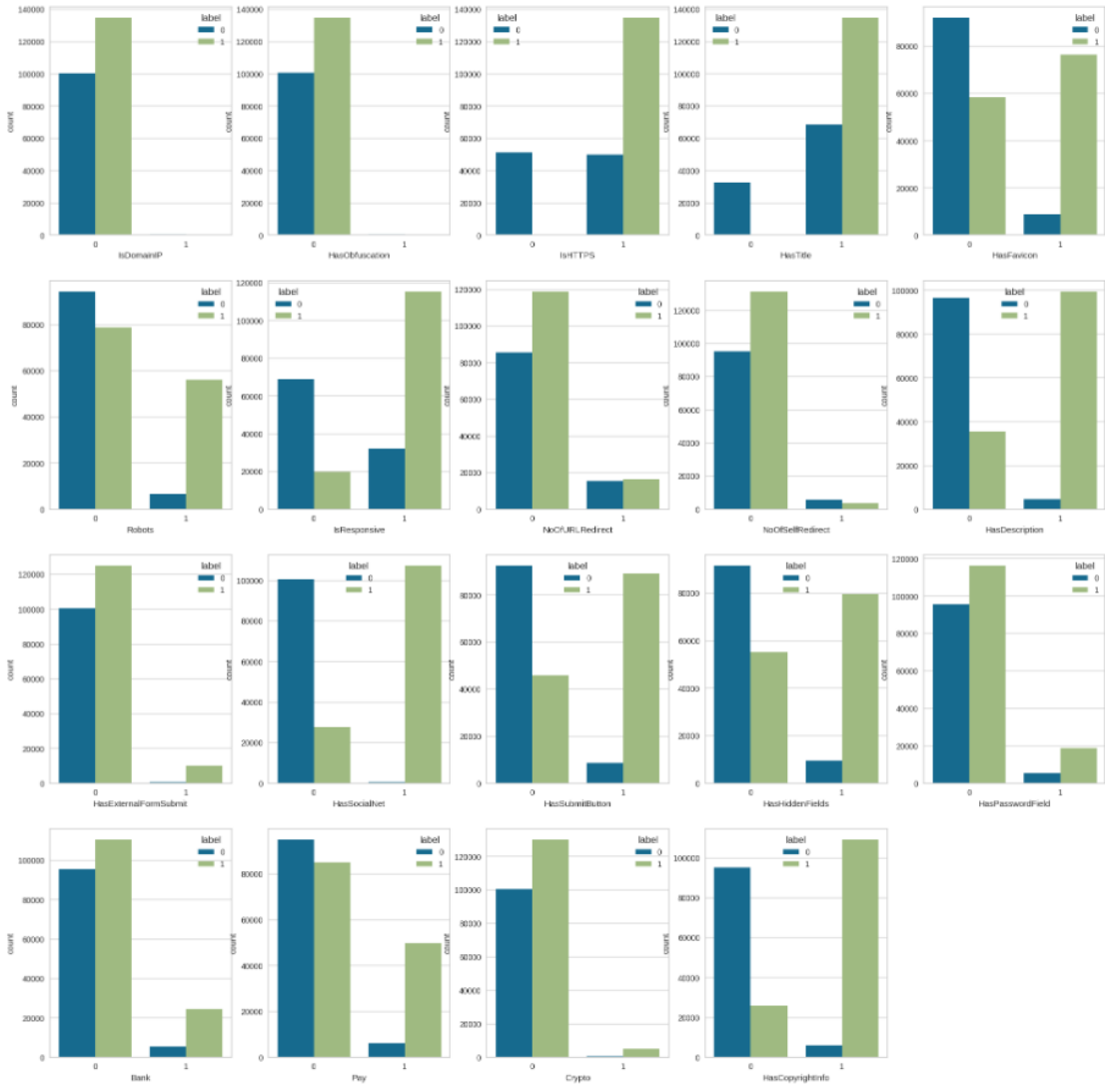
見圖二與三、編碼後合法與釣魚網站 TLD 分布以及卡方檢定



Chi-squared inpendence test	
Chi-squared statistic	27578.6415
P-value	0.0
Degrees of freedom	8

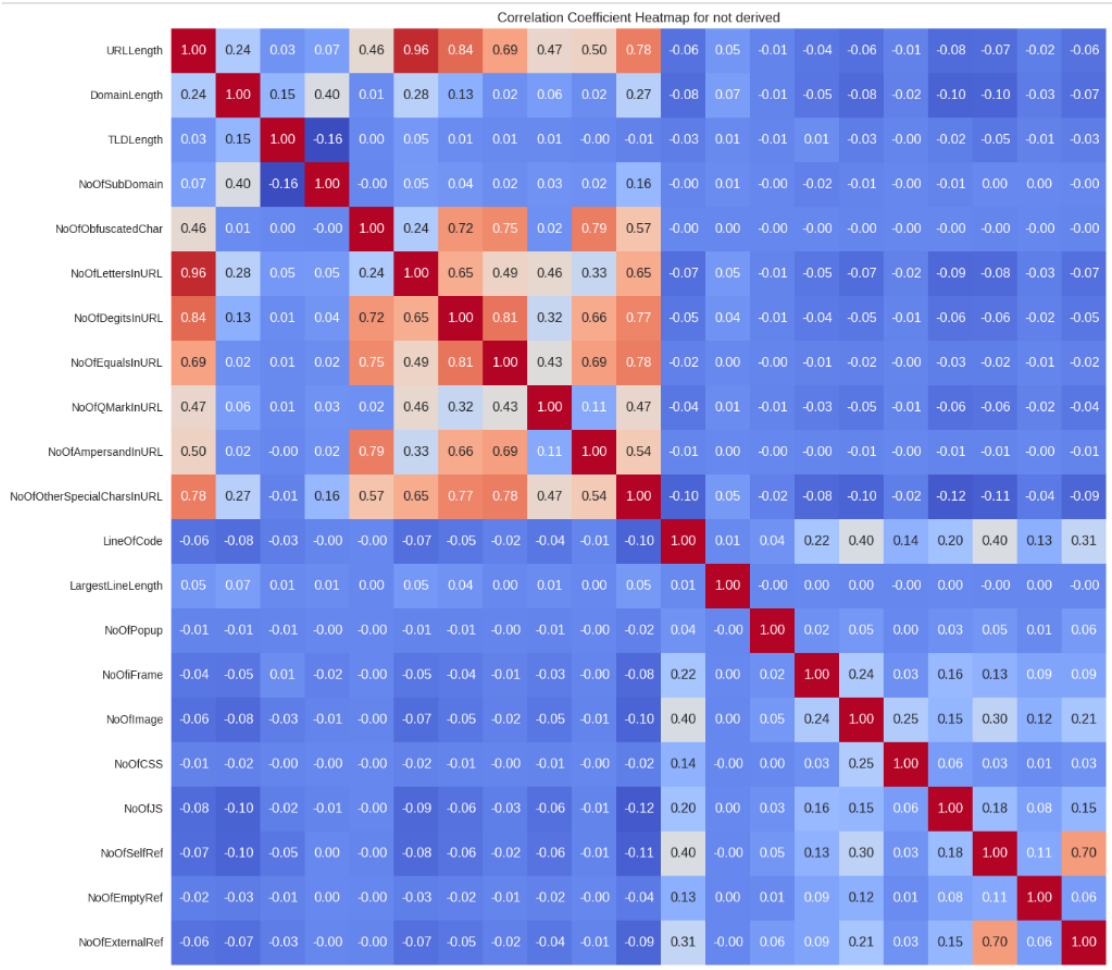
接著，觀看二元特徵的 count plot，可以觀察到，某些類別變數在預測釣魚網站時具有較高的區分能力，在建模部分可能會抓取一些特徵來建模。

見圖五、二元變數的 count plot



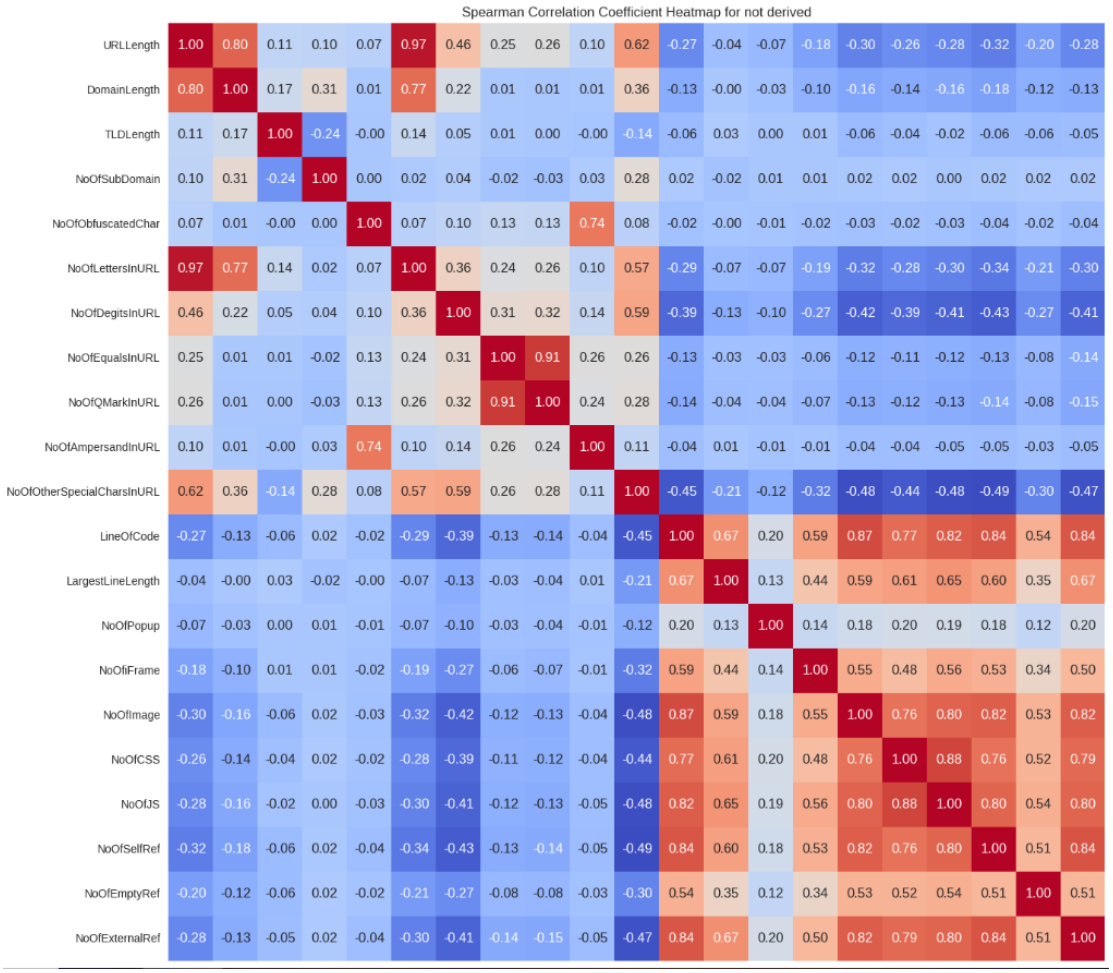
然後，我們觀看到各類數值特徵的相關係數圖，可以看出 URLLength 與 NoOfLettersInURL 有強線性關係，NO 家族關於 URL 的變數與 URLLength 彼此線性關係挺強的，可能有共線性的特徵我們只會擇一來建模。

見圖六、數值變數相關係數圖



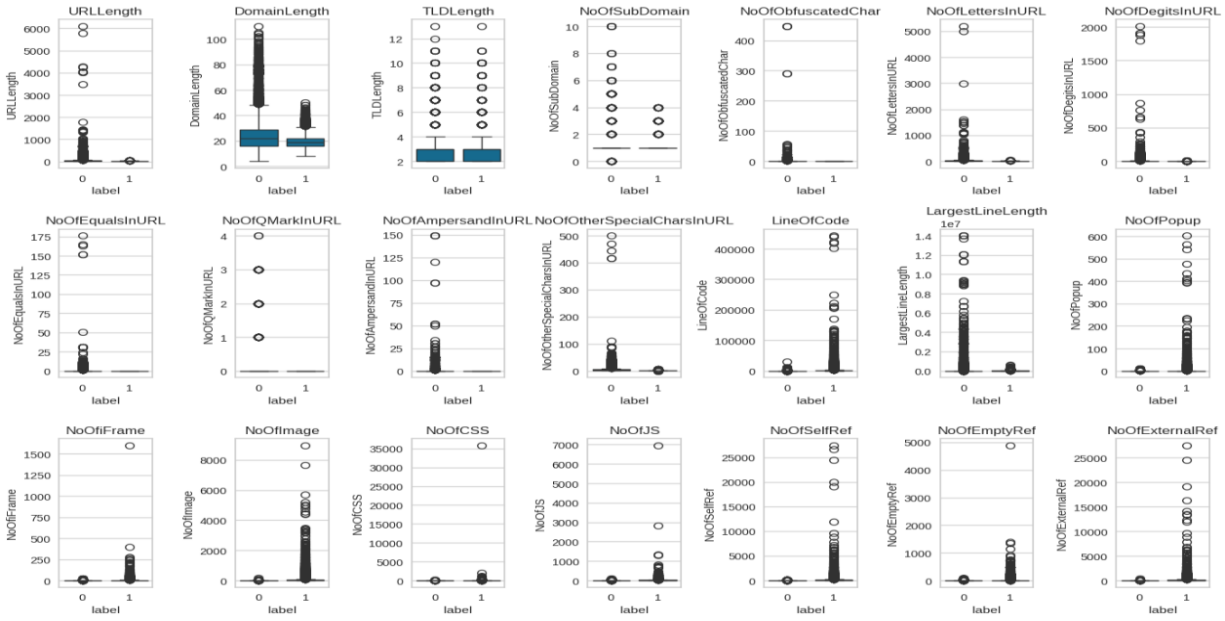
接著，我們觀察數值變數 spearman 相關係數，可以看出 URLLength 與 NoOfLettersInURL 有強非線性關係，URLLength 與 DomainLength 及其他關於 code、網站框架跟 Ref 類型的變數彼此非線性關係蠻強的，可能考慮將他們做一些變換之後建模。

見圖七、數值類變數 spearman 相關係數圖



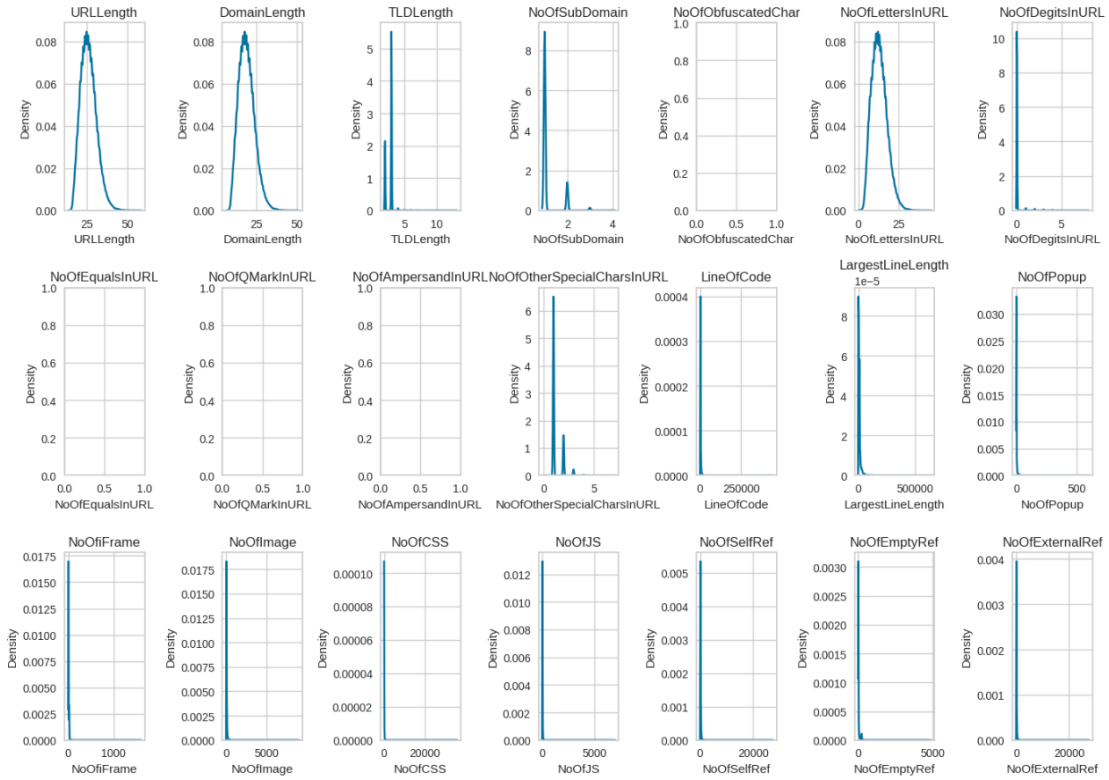
畫出以釣魚網站標籤 0 或 1 為分類的數值類變數盒鬚圖，可以發現大多數數值類變數有許多離群值，除了混淆字源數量、問號數量、連接符號(&)數量在非釣魚網站沒有，且除了 DomainLength 與 TLDLength 的其他數值類變數分配都集中在 0 附近。

見圖八、釣魚網站標籤 0 或 1 為分類的數值類變數盒鬚圖



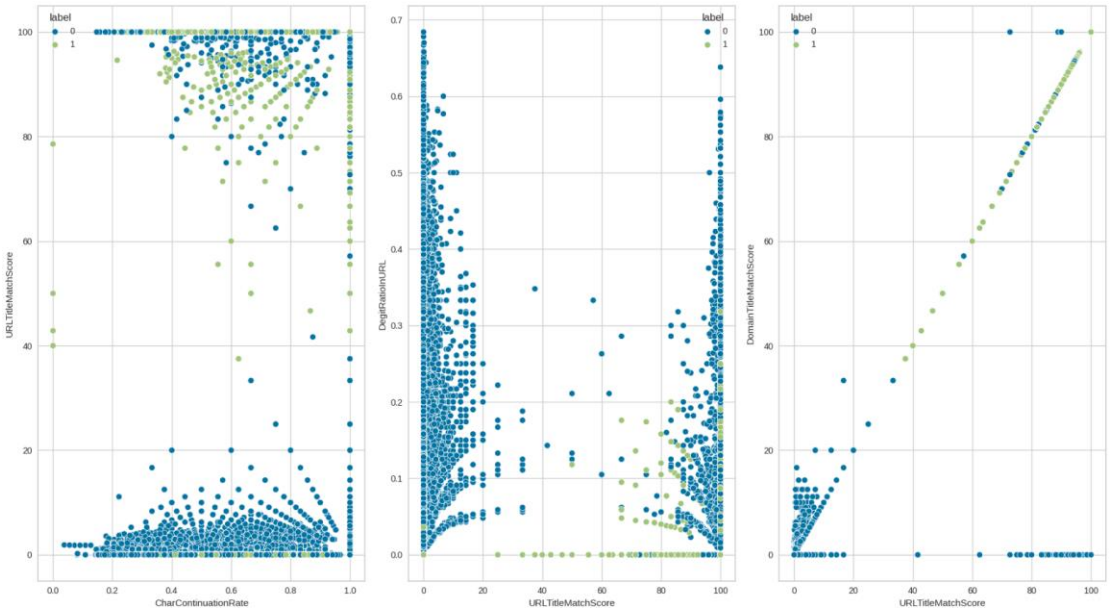
繪製數值類變數兩兩的 histogram，可以看到大多數數值類變數基本上看不出分配形式，或許標準化能減少 outlier 影響。

見圖九、數值類變數 histogram



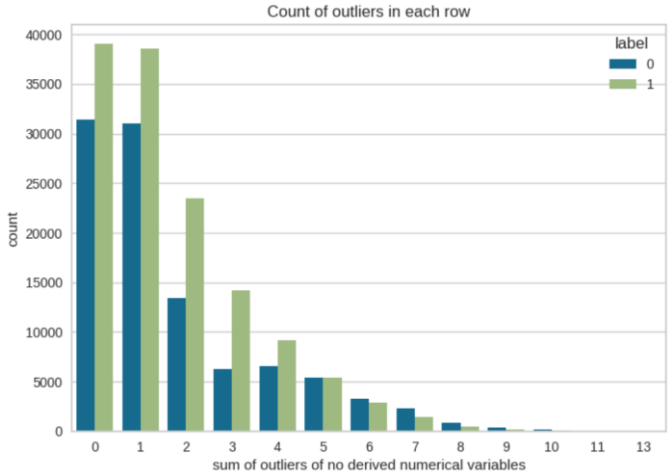
繪製 scatter plot，可以看出某些數值類變數散佈圖受到異常值的扭曲，這使得線性關係變得不那麼明顯。

見圖十、可能重要數值類變數 scatter plot



現在將有 outlier 的每一列對應的行設置成 1，否則 0。將每一列數值類變數有 outlier 的 column 相加並畫出 countplot，以標籤做分類，能看出每列 outliers 比較少的，是合法網站占比更大，最後進行 Mann-Whitney 檢定，說明釣魚網站與合法網站的數值類變數 outlier 數量分佈有顯著差異。

見圖十一與十二、每列 outliers 數量分類圖以及 Mann-Whitney 檢定結果



Mann-Whitney U statistic independence test	
Mann-Whitney U statistic	6681537497
P-value	4.12*1e-15
Degrees of freedom	15

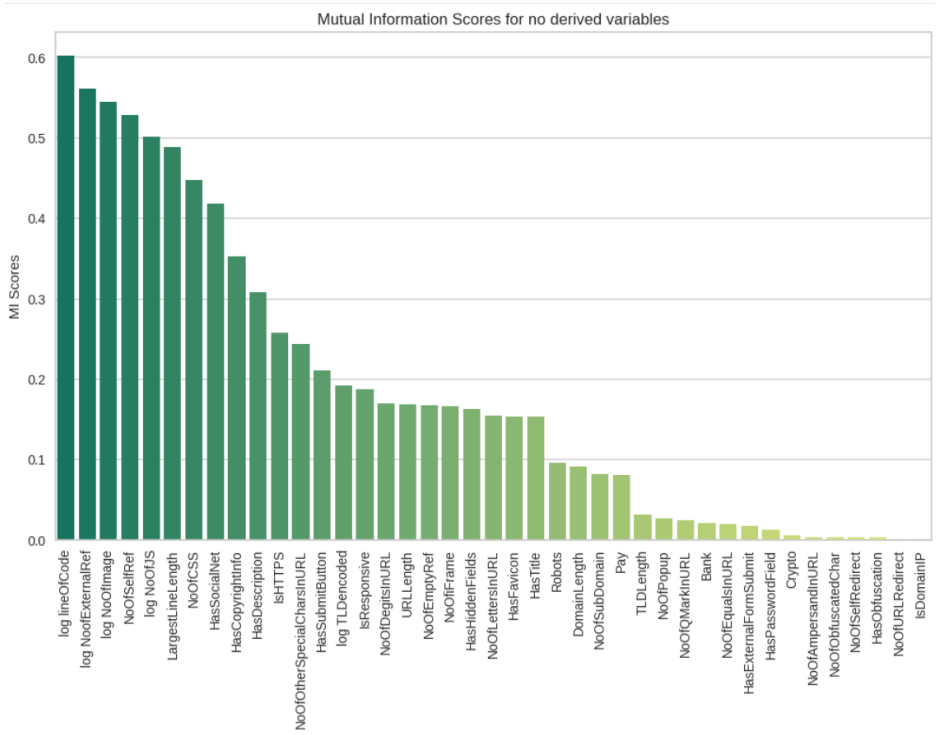
第三章 建立模型與變數選擇

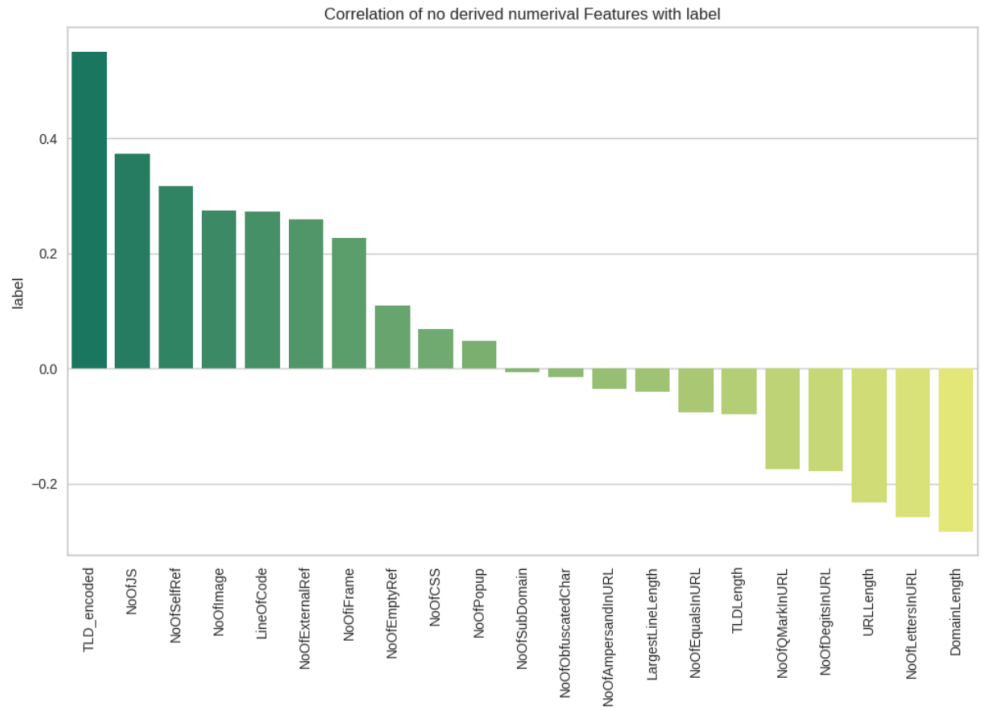
第一節 資料處理與變數選取

先將有與 label 有強非線性關係變數 Log 轉換，例如:LineOfCode，然後使用 mutual information 方法來選擇變數。

見圖十三與十四、

非人造變數 mutual information 以及人造變數 mutual information

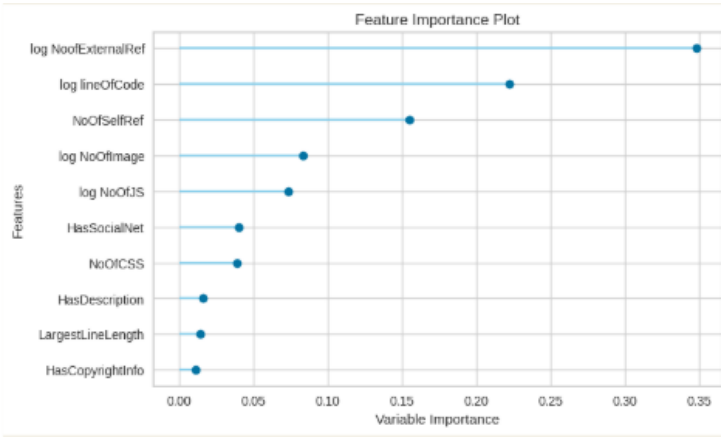




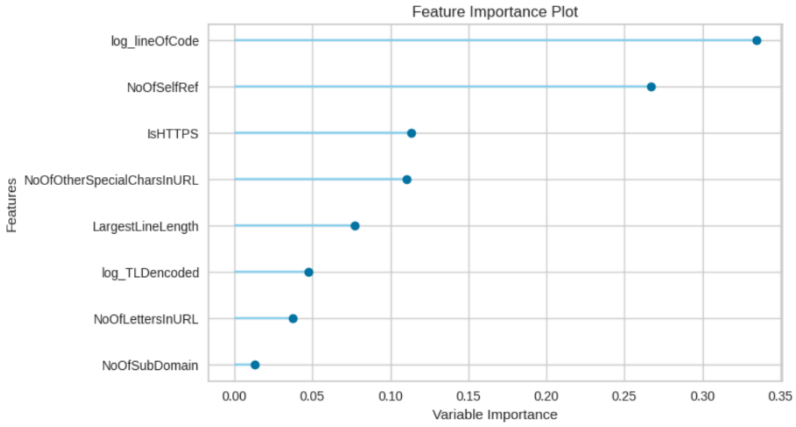
第二節 模型建立(3 種方法，(TRAIN_SIZE, TEST_SIZE)=(0.8,0.2))

我們使用 pycaret 套件標準化特徵後自動選取特徵建模並進行預測比較見下方圖十五、十六與十七、各模型的特徵重要程度圖

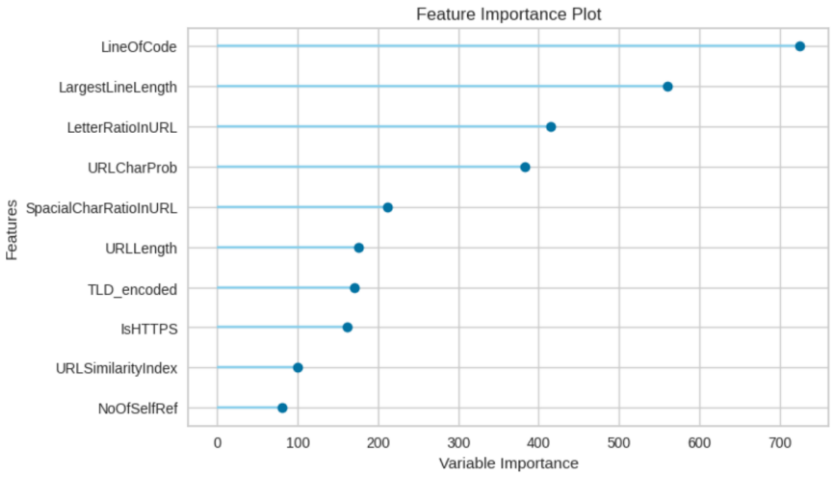
1. 使用選擇的 10 個變數建模(隨機森林)



2. 使用所有非人造變數加上 URLSimilarityIndex(隨機森林)



3.使用原資料集包括人造變數建模 (lightgbm)



第三節 小節與表格比較

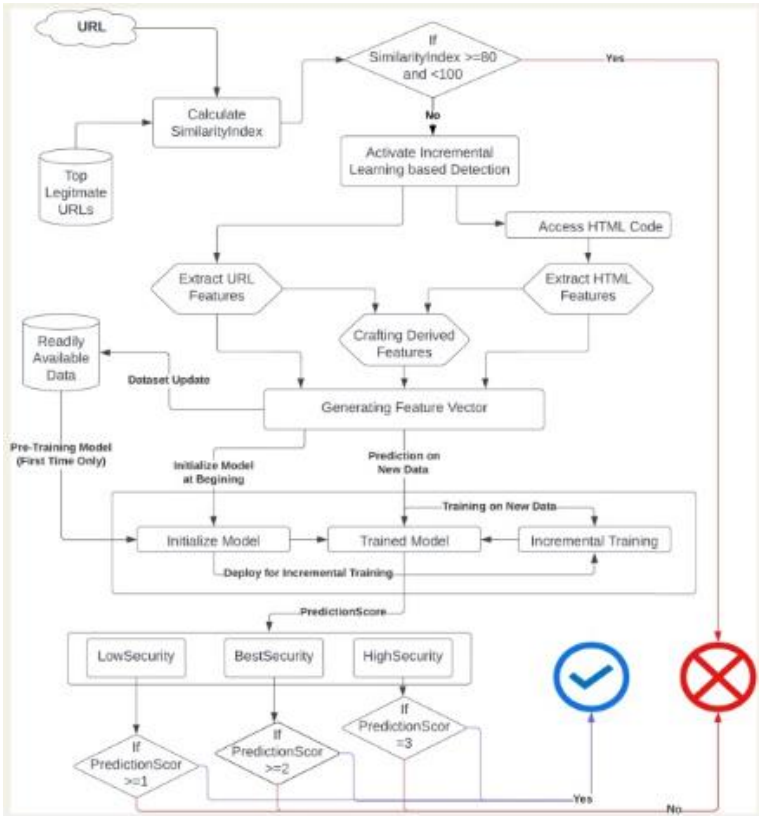
可以發現原本第一種方法所建模的交叉驗證準確度就能達到 99.7%，但第二種方法能提高 0.1%準確度，說明我們找到的 10 個特徵並不是最佳組合。且使用原資料集的人造變數可使得準確度到達 1，說明 URLSimilartyIndex 解釋了更多變數間關係。code 越長以及是 https 相關網站有很高可能性是合法網站。URL 相似性指標確實與論文內容說得一樣是關鍵變數。

見圖十八、三種模型表格比較

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	CV	備註
1.隨機森林分類器	0.9967	0.9999	0.9977	0.9966	0.9971	0.9933	0.9933	0.997	
2.隨機森林分類器	0.9996	1.0000	0.9996	0.9996	0.9996	0.9991	0.9991	0.999	多了 IsHTTPS
3.LightGBM	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.000	

第四章 論文實作

見圖十九、論文內釣魚網站處理流程圖



見圖二十、將原資料集每 2000 筆 data 切成一個 batch，使用 naive bayes bernoulli model 所做出增量學習實作結果

第五章 總結

本研究通過對釣魚網站檢測的深入探討，驗證了 URL 相似性指標（URL Similarity Index）的有效性。結果顯示，該指標不僅能顯著提高模型的準確性，還可以更全面地捕捉釣魚網站與合法網站之間的潛在差異。此外，研究還發現 HTTPS 的使用和程式碼的長度與合法網站的可能性密切相關，說明這些特徵在區分釣魚網站和合法網站時具有重要的參考價值。 在模型性能方面，本研究的模型在靜態資料集上表現卓越，最佳模型達到了 100% 的準確度，展現了研究框架的理論有效性。然而，當面臨快速變化的網路環境時，模型的適應性不足成為一個挑戰。靜態學習框架無法及時處理不斷增長的動態數據，因此每次模型的重新訓練既耗時又資源密集，無法滿足實際應用需求。 針對上述挑戰，增量學習被認為是未來應對此問題的關鍵技術之一。與傳統的靜態學習方法不同，增量學習能夠在新的數據到達時進行局部更新，而無需重新訓練整個模型。這不僅提高了模型的運行效率，還能確保模型在動態網路環境中保持高準確性。增量學習還適合於處理超大數據集和資源受限場景，尤其是在需要頻繁更新數據的實務應用中。 未來，研究可以進一步探討增量學習框架的優化，並結合深度學習技術來提升檢測性能。此外，還可以嘗試應用其他高效的數據處理技術，例如在線學習和流數據處理，以進一步增強系統的靈活性和實用性。同時，考慮到釣魚網站攻擊手法的多樣性，未來模型的開發可以融入更多語義分析、網頁結構特徵等多維度特徵，以提高釣魚網站檢測的全面性和穩健性。 綜上所述，本研究不僅為釣魚網站檢測提供了有效的框架，還為未來的網路安全技術發展提供了重要啟示。模型的高準確度和增量學習的潛在應用將在實務中發揮關鍵作用，特別是在當前網路安全威脅不斷增長的背景下，為實現更智能、更高效的網路安全防護奠定基礎。

Reference

[1]PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning
資料集網址:
https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+dataset