ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
HIGH PERFORMANCE COMPUTING LABORATORY - BIG DATA CLUB

BDC ASSIGNMENT 3 - PROJECT PITCHING

PHISHING WEB DETECTION USING MACHINE LEARNING METHODS

GVHD: Hoàng Lê Hải Thanh
Nhóm: 03
SV thực hiện: Phạm Đại Hoàng An
Trần Hoàng Công Toại
Nguyễn Trọng Anh
Đổng Hoàng Sơn

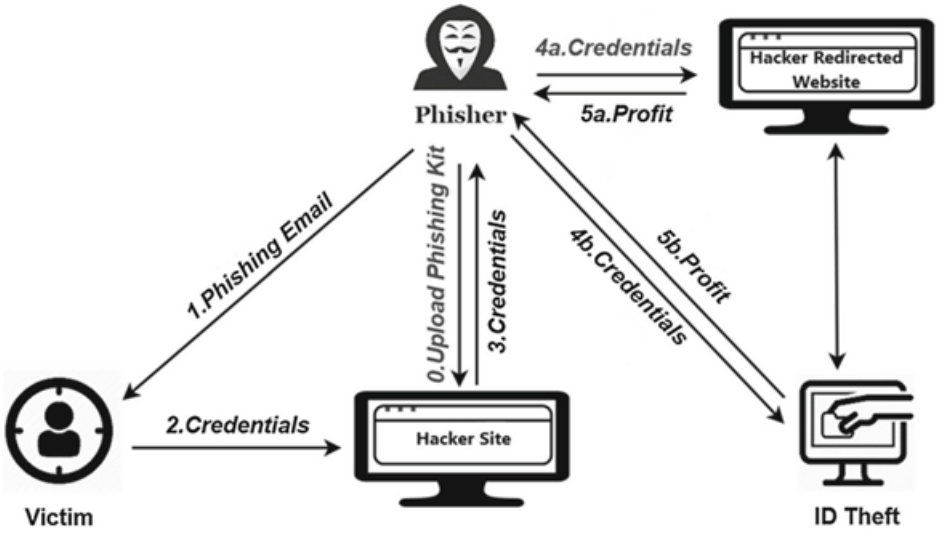
Tp. Hồ Chí Minh, Tháng 08/2022

Mục lục

Danh sách hình vẽ

I. Giới thiệu đề tài

Trong những năm trở lại đây, tấn công lừa đảo (phishing attack) là một trong những cuộc tấn công nguy hiểm và trọng yếu mà những người dùng internet, chính phủ,... Ở các cuộc tấn công này, những kẻ tấn công sẽ dùng các kỹ thuật khác nhau để nhằm chiếm đoạt những thông tin của người dùng (tài khoản và mật khẩu ngân hàng, thông tin tùy thân...). Các kỹ thuật mà chúng thường hay sử dụng nhất là thư điện tử và trang web giả mạo (spoofed emails and fake websites). Những kẻ này sẽ tạo thành một trang web/email có hình thức và nội dung gần như là một bản sao chép từ các trang web/email chính chủ uy tín. Và khi người dùng đưa các thông tin về tài khoản, mật khẩu, thông tin thẻ ngân hàng,... chúng sẽ có được toàn bộ thông tin này. Trong năm 2018, theo nhóm Công tác Chống lừa đảo (Anti-Phishing Working Group – APWG), có khoảng 51,401 trang web lừa đảo. Một báo cáo khác của RSA chỉ ra rằng các tổ chức và hiệp hội trên thế giới chịu thiệt hại lên đến 9 tỷ đô bởi các cuộc tấn công lừa đảo trong năm 2016. Với sự phát triển mạnh mẽ của lĩnh vực máy tính nói chung, ta ngày càng có nhiều công cụ và biện pháp để tránh khỏi các cuộc tấn công đó, tuy nhiên ta cần hiểu cơ bản về các mà các cuộc tấn công lừa đảo hoạt động. Hình bên dưới sẽ miêu tả rõ điều đó:



Hình 1: Quy trình lừa đảo thông qua email và trang web giả mạo

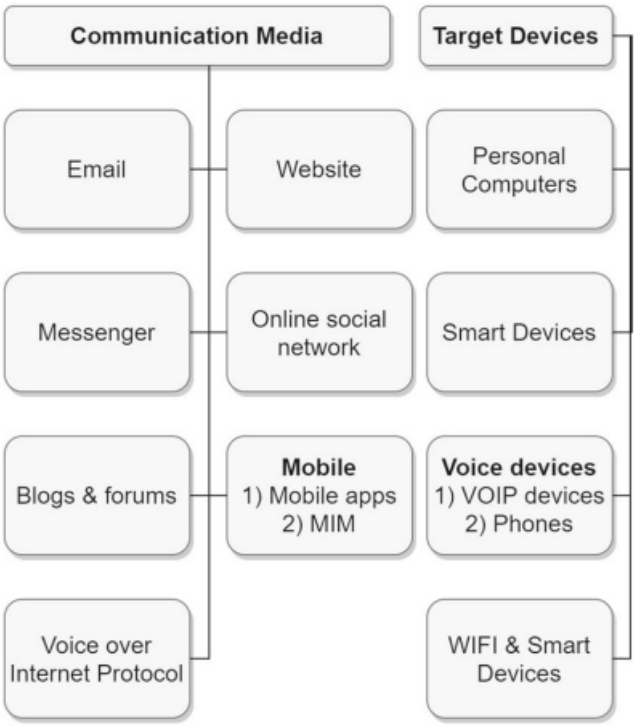
1. Các kẻ lừa đảo sẽ gửi email cho nạn nhân, các email này có hình thức và nội dung tương tự các email họ thương được nhận (quảng cáo, thông báo giảm giá,...), nhưng tại đây, những kẻ ấy sẽ chèn URL hướng về các website giả đã được dựng sẵn.

2. Khi nạn nhân đã tin tưởng trang web, họ sẽ nhập các thông tin mà các kẻ lừa đảo

mong muốn.

3. Sau đó, các thông tin sẽ được gửi về cho máy chủ của những kẻ lừa đảo. Từ đây, họ đã có những thông tin cần thiết để chiếm đoạt quyền sử dụng tài khoản, chiếm đoạt tài khoản ngân hàng của các nạn nhân.

Dưới đây là tổng hợp các phương tiện mà những kẻ lừa đảo thường sử dụng:



Hình 2: Các phương tiện mà những kẻ lừa đảo thường tiếp cận

Người dùng cá nhân được cho là những "con mồi béo bở" cho những kẻ lừa đảo, bởi lẽ:

- Đa số người dùng đều chưa có những thông tin cơ bản về Uniform Resource Locator (URLs).
- Địa chỉ thật của các trang web giả mạo đã bị giấu hoặc ẩn đi, dẫn đến họ không biết được họ đang bị điều hướng đến một trang web lừa đảo.
- Họ chưa có kinh nghiệm và suy nghĩ về một trang web đáng tin cậy.
- Có nhữn người dùng không thể phân biệt được trang web chính thống và trang web lừa đảo.

Các trang web lừa đảo mà những kẻ này tạo ra, thường có những đặc điểm sau:

- Có vẻ bề ngoài và các chứng năng gần như giống hết với trang web chính gốc.
- Có đường dẫn URL kì lạ, không có tên miền đáng tin.
- Một số trang web lừa đảo vẫn sử dụng giao diện cũ của những trang web chính thức.

Tuy nhiên, các tên lừa đảo ngày càng tinh vi và được trang bị nhiều kiến thức hơn, dẫn đến nhiều khó khăn hơn trong việc xác định các trang web lừa đảo. Chúng ngày càng nâng cấp thủ thuật như:

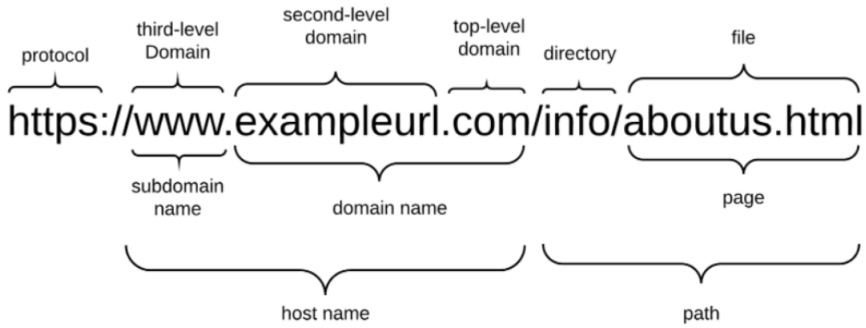
- Sử dụng lỗi bảo mật của chính trang web chính chủ để chèn chức năng lấy thông tin của người dùng trước khi thông tin được nhập vào trang web chính thức.
- Sử dụng URL giả, ban đầu đó là URL chính thống, nhưng khi ấn vào, ta sẽ được điều hướng sang một trang web giả đã được định trước.

Chính bởi các hành vị lừa đảo ngày càng tinh vi và khó nhận biết, và cũng để bảo vệ cho nhiều người dùng chưa có được những kiến thức cần thiết, sự ra đời của một công cụ giúp bảo vệ người dùng, tăng cường bảo mật và bảo đảm sự an toàn thông tin là cấp thiết. Tuy nhiên hiện nay, đa số những ứng dụng như thế đều chưa thật sự đạt hiệu quả cao nhất, do chúng thường là nhận được báo cáo về độ đáng tin cậy của người dùng, và từ đó cảnh báo lại cho những người dùng sau. Mô hình này sẽ không hoàn toàn đúng, khi mà những kẻ xấu có thể đánh giá tốt cho trang web của chúng, và người dùng hoàn toàn vẫn có thể đánh giá sai về trang web ấy. Chính vì thế, việc ứng dụng mô hình tự học - Machine Learning và Deep Learning - sẽ giúp tăng độ chính xác, cũng như sẽ ngày càng tự hoàn thiện bản thân.

II. Mô tả dữ liệu

1. Tổng quan về Dataset

Việc lừa đảo qua mạng đang dần dần được phổ biến với những cách thức tinh vi khác nhau, với việc giả dạng cho giống một trang web thật để đánh lừa người dùng. Đây là một hình thức lừa đảo trong đó kẻ tấn công cố gắng tìm hiểu thông tin nhạy cảm như thông tin xác thực đăng nhập hoặc thông tin tài khoản bằng cách giả vờ là một tổ chức hoặc cá nhân có uy tín thông qua email hoặc các phương tiện liên lạc khác.
URL của các trang web lừa đảo có thể rất giống với các trang web thật đối với mắt người, nhưng chúng khác nhau về IP.



Hình 3: Tổng quan về một URL

- Domain name : đây là phần bắt buộc vì nó phải được đăng ký với tên miền.
- Subdomain name và Path : hoàn toàn có thể bị những kẻ lừa đảo kiểm soát và đánh lừa người dùng.

2. Các kiểu dữ liệu trong URL

Các đặc điểm của một URL bao gồm :

1. **Độ dài của URL** : Những người lừa đảo có thể che giấu các thành phần đáng ngờ của website giả mạo bằng cách sử dung URL có độ dài lớn.
2. **Tiền tố và hậu tố** : được phân các bởi ký tự ’-’ trong URL. Những website giả mạo có thể sử dụng tiền tố, hậu tố để đánh lừa người dùng rằng họ đang sử dụng website uy tín.
3. **Số lượng "@" xuất hiện trong URL** : truy cập vào URL khiến cho người dùng bỏ qua thông tin đằng trước "@", và địa chỉ website thực thế luôn dựa trên ký tự "@"
4. **Số lượng miền phụ trong URL (subdomain)** : những người lừa đảo có thể sử dụng nhiều miền phụ để che giấu hành vi đáng ngờ trong URL.
5. **Thời gian tên miền tồn tại của website** : các website giả mạo thường có đặc điểm chung là thời gian tên miền tồn tại khá ngắn.
6. **Registration time** : Thời gian đăng ký tên miền.
7. **SFH** : Kiểm tra Server Form Handler (SFH) chứa chuỗi rỗng, chuỗi "about:blank" hay chứa tên miền khác.
8. **Abnormal URL** : Kiểm tra tên miền có trong URL hay không,

9. **Số lượng website được tìm thấy thông qua dữ liệu của WHOIS** : đây là công cụ giúp kiểm tra các thông tin của tên miền URL.

10. **Độ tin tưởng của website** : theo tiêu chuẩn của `https://www.domcop.com/openpagerank/?fbclid=IwAR2rbVhz81hDiDbKJKSDvaW2S6ik6wlbkZKP1_3yinHjF3vjnI9Ndoz3e9g`

11. **Số lượng website được tìm thấy trên Google index** : các website lừa đảo thường không xuất hiện trong công cụ Google Index.

12. **Địa chỉ IP** : Địa chỉ IP được sử dụng trong URL thay vì trong tên miền. Ví dụ như `http://217.102.21.305/sample.html`

13. **URL rút gọn** : Cung cấp các URL được rút gọn có độ dài ngắn hơn URL gốc. Ví dụ `http://sample.web.vn/` được thu gọn thành `bit.ly/Gdaj123`

14. **Chuyển hướng bằng dấu gạch chéo kép "//"** : Chỉ ra cho người dùng biết họ sẽ được chuyển hướng đến một website khác.

15. **Biểu tượng (favicon) của URL** : Biểu tượng của một URL là hình ảnh ảo được kết nối tới trang web cụ thể. Nếu một favicon được load từ miền khác với favicon ở trên thanh địa chỉ thì đó chắc chắn là website lừa đảo.

16. **Thẻ HTTPS** : Sử dụng các các thẻ http hay https trong URL.

17. **Request URL** : Là một URL mà người dùng sẽ sử dụng để truy cập vào ứng dụng web.

18. **Email cá nhân** : Các kẻ tấn công mạng có thể lấy thông tin cá nhân của người dùng thông qua chuyển hướng đến Email của người dùng.

19. **Số lượng Redirect của website** :Tính số lần chuyển trang của web.

20. **Status bar** : Có onMouseOver thay đổi thanh trạng thái không.

21. **Nhấp chuột** : Kiểm tra nhấp chuột có bị vô hiệu hóa hay không.

22. **Pop-up Window** : Kiểm tra Pop-up của Window có cảnh báo khi truy cập trang web không.

23. **IFrame** : Kiểm tra thẻ iframe để hiển thị trang web bổ sung có được sử dụng hay không.

24. **Web Traffic** : Độ nổi tiếng của website.

25. **Links pointing to page** : Số liên kết trỏ tới website.

Bằng việc sử dụng kỹ thuật filter để lấy các thông tin cần thiết từ dữ liệu tổng như length_url, prefix_suffix,... và trực quan hóa lên biểu đồ thông qua python, từ đó đưa ra được đặc điểm cụ thể của từng feature trong legit website và phishing website.

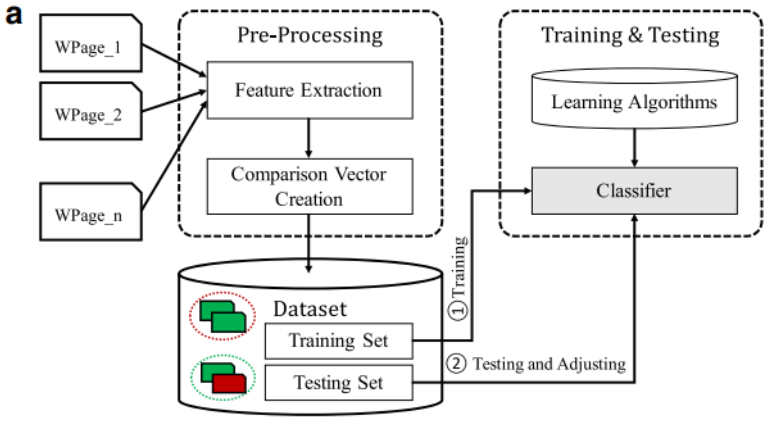
Từ các biểu đồ đã được trực quan trong mỗi trường hợp, ta đưa ra được các kết luận về đặc điểm của một phishing website. Từ đó tạo nên nguồn dữ liệu để train và test trong quá trình train model sau này. Nếu một website có nhiều yếu tố của một website lừa đảo thì ứng dụng sẽ thông báo đây là website lừa đảo.

Việc loại bỏ các giá trị không cần thiết (các giá trị không giúp chúng ta phân biện phishing website) và số hóa dữ liệu sang số liệu mình sẽ dùng để train model giúp cho tập train và test của model trở nên gọn gàng và vừa đủ thông tin, tiếp kiệm thời gian trong việc xử lý dữ liệu vô giá trị.

III. Phương pháp phát hiện trang Web lừa đảo

Trong phần này, nhóm sẽ giới thiệu và mô tả giải pháp mà nhóm đã sử dụng. Để có thể phân loại được các trang Web giả mạo, nhóm đã phân ra hai giai đoạn bao gồm: *Training mô hình phân loại* và *Phát hiện trang web giả mạo dựa trên mô hình phân loại*.

1. Training mô hình phân loại

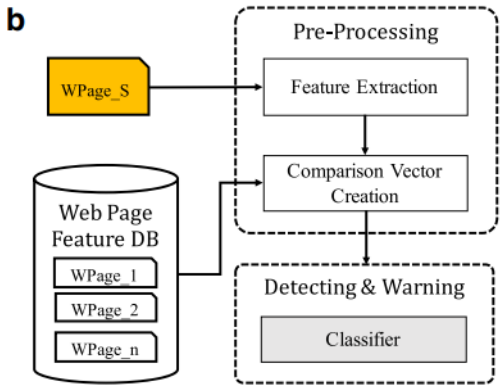


Hình 4: Quy trình huấn luyện mô hình

Như đã thấy ở hình trên, giai đoạn này sẽ bao gồm giai đoạn tiền xử lý (pre-processing) và giai đoạn đào tạo (training). Ở giai đoạn tiền xử lý ta cần chuẩn bị tập dữ liệu là các địa chỉ **URL** với các đặc tính được chọn từ trước. Các địa chỉ trang Web trong tập dữ liệu phải được đánh dấu là giả mạo hay không để dễ dàng xử lý. Từ các tính năng đã được chọn lọc, ta trực quan hóa dữ liệu giúp xác định được đặc tính cụ

thể của một trang Web giả mạo và một trang Web hợp lệ khác nhau như thế nào, tóm tắt các tính năng tương đồng đó và lọc lại dữ liệu dựa trên các tính năng đó. Sau khi lọc dữ liệu và phân loại dựa trên đặc điểm của từng tính năng, tập dữ liệu sẽ được phân ra thành tập dữ liệu để training và tập dữ liệu dùng để test độ chính xác. Với tập dữ liệu dùng để training, ta đưa vào một mô hình phân loại cụ thể để huấn luyện mô hình đó, sau đó dùng tập test để kiểm tra độ chính xác của mô hình.

2. Phát hiện trang web giả mạo dựa trên mô hình phân loại



Hình 5: Quy trình phát hiện trang Web lừa đảo

Sau quá trình huấn luyện mô hình, chúng ta có thể sử dụng để phân loại các trang Web từ đó phát hiện trang Web giả mạo. Khi người dùng sử dụng một trang Web, trang Web đó với mã URL sẽ được xử lý dữ liệu thông qua các đặc tính được chọn, các dữ liệu từ thanh địa chỉ đó sẽ được đưa qua mô hình mà ta đã chọn để phân loại xem nó có phải là giả mạo hay không. Khi phân loại xong, URL đó sẽ được dan nhãn và từ đó hệ thống dựa vào nhãn đó mà cảnh báo đến cho người dùng.

IV. Các thuật toán học máy sử dụng

Để đánh giá được độ tin cậy cũng như chính xác của các giải thuật sử dụng, áp dụng các thông số sau:

- $Precision = \frac{TP}{TP+FP}$: cho biết tỉ lệ dự đoán đúng trang web giả mạo thật sự trong tổng số trang web được dự đoán là giả mạo. Giá trị này càng cao thì chứng tỏ tỉ lệ dự đoán chính xác trang web giả mạo của mô hình càng cao.
- $Recall = \frac{TP}{TP+FN}$: cho biết tỉ lệ số trang web được dự đoán đúng là giả mạo trong tổng số trang web giả mạo thật sự. Giá trị này càng cao thì chứng tỏ tỉ lệ mô hình dự đoán được trang web giả mạo càng cao.
- $F1-score = \frac{2*Recall*Precision}{Recall+Precision}$: trung bình điều hòa của *Precision* và *Recall*.

– $Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$: độ chính xác của mô hình, tức là tỉ lệ số lượng dự đoán đúng so với tổng số trang web.

Trong đó:

– TP - True Positive là số lượng class 1 được dự đoán là class 1.

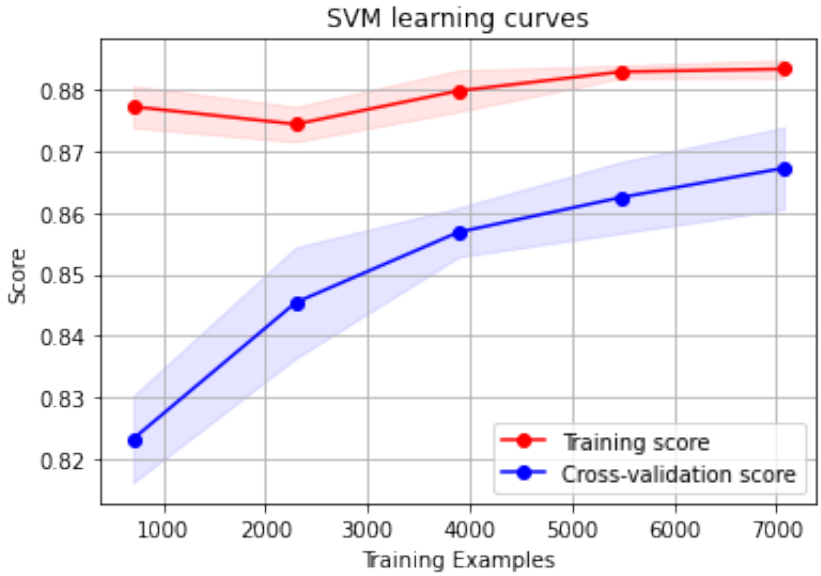
– FP - False Positive là số lượng class -1 được dự đoán là class 1.

– TN - True Negative là số lượng class -1 được dự đoán là class -1.

– FN - False Negative là số lượng class 1 được dự đoán là class -1.

1. Support Vector Machine

Support Vector Machine (SVM) là một thuật toán học có giám sát, được sử dụng chủ yếu cho bài toán phân loại. Trong thuật toán này, ta biểu diễn đồ thị dữ liệu là các điểm dữ liệu *n* chiều, sau đó ta sẽ thực hiện tìm siêu phẳng (hyper-plane) tốt nhất phân chia các lớp.



Hình 6: Huấn luyện mô hình bằng giải thuật SVM

Thực hiện training model với giải thuật SVM, ta thấy được *training score* và *cross-validation score* có chiều hướng tăng lên khi tập dữ liệu của mô hình tăng lên. Vì vậy khi lượng dữ liệu càng nhiều thì độ chính xác càng cao, phù hợp với việc phát triển ứng dụng sau này.

Ưu điểm

SVM là một kỹ thuật phân lớp khá phổ biến, nó thể hiện được nhiều ưu điểm trong số đó có việc tính toán hiệu quả trên các tập dữ liệu lớn, ngoài ra còn có một số các ưu điểm chính sau:

- Xử lý trên không gian số chiều cao: SVM là một công cụ tính toán hiệu quả trong không gian nhiều chiều, trong đó đặc biệt áp dụng cho các bài toán phân loại văn bản và phân tích quan điểm nơi số chiều có thể cực kỳ lớn.

- Tiết kiệm bộ nhớ: do chỉ có một tập hợp con của các điểm được sử dụng trong quá trình huấn luyện và ra quyết định thực tế cho các điểm dữ liệu mới nên chỉ có những điểm cần thiết mới được lưu trữ trong bộ nhớ.

- Tính linh hoạt: khả năng áp dụng của Soft Margin SVM và Kernel SVM cho phép linh động giữa các phương pháp tuyến tính và phi tuyến tính từ đó khiến cho hiệu suất phân loại lớn hơn.

Nhược điểm

- Bài toán số chiều cao: trong trường hợp số lượng thuộc tính của tập dữ liệu lớn hơn rất nhiều so với số lượng dữ liệu thì SVM cho kết quả không tốt lắm.

- Chưa thể hiện rõ tính xác suất: việc phân lớp của SVM chỉ là việc cố gắng tách các đối tượng vào hai lớp được phân tách bởi siêu phẳng SVM. Điều này chưa giải thích được xác suất xuất hiện của một thành viên trong một nhóm là như thế nào. Tuy nhiên hiệu quả của việc phân lớp có thể được xác định dựa vào khái niệm *margin* từ điểm dữ liệu mới đến siêu phẳng phân lớp.

2. Logistic Regression

Hồi quy Logistic (Logistic Regression) là một thuật toán học có giám sát, tương tự như thuật toán hồi quy tuyến tính, chỉ khác ở chỗ từ đầu ra của hàm tuyến tính, ta đưa vào hàm Sigmoid để tìm ra xác suất của dữ liệu thuộc nhãn lớp. Vì vậy, nó được sử dụng cho bài toán phân loại.

Hình 7: Huấn luyện mô hình bằng giải thuật Logistic Regression

Thực hiện training model với giải thuật Logistic Regression, ta thấy được *training score* và *cross-validation score* có xu hướng hội tụ về một giá trị thấp hơn so với ban đầu. Vì vậy khi càng nhiều dữ liệu, mô hình sẽ không đạt được hiệu quả như mong muốn.

Ưu điểm

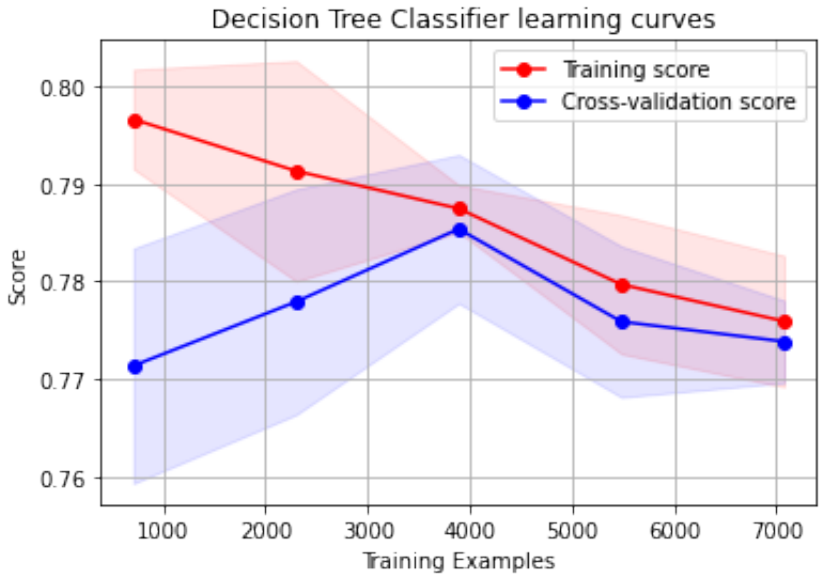
- Mô hình dễ hiểu.
- Phân loại khá tốt trong trường hợp dữ liệu gần với linearly separable.
- Phân lớp nhanh.

Nhược điểm

- Không phù hợp với các loại dữ liệu phi tuyến.
- Yêu cầu các điểm dữ liệu được tạo ra một cách độc lập với nhau.

3. Decision Tree

Cây quyết định (Decision Tree) là một thuật toán học có giảm sát, nó là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượng có thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary) , Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.



Hình 8: Huấn luyện mô hình bằng giải thuật Decision Tree

Thực hiện training model với giải thuật Decision Tree, ta thấy được *training score* và *cross-validation score* có xu hướng hội tụ về một giá trị thấp hơn so với ban đầu giống với mô hình Logistic Regression nên khi lượng dữ liệu càng lớn thì độ hiệu quả sẽ không như mong muốn.

Ưu điểm

- Mô hình dễ hiểu và dễ giải thích.
- Cần ít dữ liệu để huấn luyện.
- Có thể sử dụng cho cả dữ liệu số và dữ liệu phân loại.
- Xây dựng nhanh.
- Phân lớp nhanh.

Nhược điểm

- Không đảm bảo xây dựng được cây tối ưu.
- Có thể overfitting.
- Thường ưu tiên thuộc tính có nhiều giá trị.

4. Random Forest

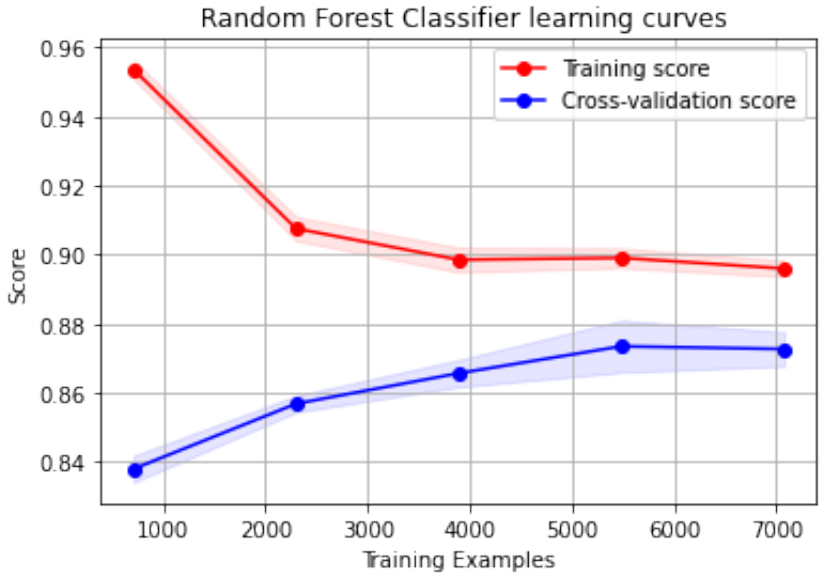
Rừng ngẫu nhiên (Random Forest) là thuật toán học có giám sát, nó được xây dựng dựa trên thuật toán Decision Tree. Ở đây thay vì chỉ tạo ra một cây quyết định thì

Random Forest sẽ tạo ra nhiều cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên. Vì vậy, kết quả dự đoán sẽ được chọn bằng cách bỏ phiếu từ kết quả dự đoán của mỗi cây.

Random Forest gồm nhiều cây quyết định, mỗi cây quyết định đều có những yếu tố ngẫu nhiên:

- Lấy ngẫu nhiên dữ liệu để xây dựng cây quyết định.
- Lấy ngẫu nhiên các thuộc tính để xây dựng cây quyết định.

Mỗi cây quyết định không được xây dựng từ toàn bộ tập dữ liệu cũng như không dùng tất cả các thuộc tính nên mỗi cây có thể dự đoán không tốt, khi đó mỗi cây quyết định không bị overfitting mà có thể bị underfitting. Tuy nhiên, kết quả cuối cùng của Random Forest lại tổng hợp từ nhiều cây quyết định, thế nên thông tin các các cây sẽ bổ sung cho nhau, và kết quả thực nghiệm cho thấy mô hình có kết quả dự đoán tốt hơn so với 1 cây quyết định.



Hình 9: Huấn luyện mô hình bằng giải thuật Random Forest

Mặc dù có độ chính xác cao hơn, nhưng cũng giống với giải thuật Logistic Regression và Decision Tree, ta thấy được *training score* và *cross-validation score* cũng có xu hướng hội tụ về một giá trị thấp hơn nên sẽ không ưu tiên hàng đầu khi lựa chọn thuật toán này.

5. Áp dụng các mô hình vào bài toán phân loại Web lừa đảo

Sử dụng tập dữ liệu gồm 11054 mẫu, trong đó 6157 mẫu có class 1 và 4897 mẫu có class -1 (class 1 là web lừa đảo và -1 là không).

– Tập train: 8843 mẫu (80% tập dữ liệu ban đầu), trong đó 4937 mẫu có class 1 và 3906 mẫu có class -1.

– Tập test: 2211 mẫu (20% tập dữ liệu ban đầu), trong đó 1220 mẫu có class 1 và 991 mẫu có class -1.

Sử dụng các mô hình trên trong thư viện *scikit-learn* để huấn luyện thì thu được kết quả như sau:

	Precision	Recall	F1-score	Accuracy
SVM	0.89	0.89	0.89	0.88
Logistic Regression	0.83	0.82	0.82	0.81
Decision Tree	0.74	0.94	0.83	0.78
Random Forest	0.90	0.88	0.89	0.88

Kết luận: Dựa vào mô hình phân loại khác nhau, dễ dàng thấy được thuật toán Support Vector Machine phù hợp với việc phát triển cũng như ứng dụng trong việc phát hiện trang Web giả mạo. Với dữ liệu được mở rộng thì độ chính xác của mô hình SVM sẽ được tăng cao, từ đó sẽ dễ dàng phát hiện hơn.

V. Giá trị khoa học và hướng phát triển đề tài

1. Giá trị khoa học

Thời đại công nghệ ngày càng phát triển và phổ biến ở mọi linh vực kinh tế, chính trị, giao thông, giải trí,... do đó hầu hết các thông tin, tài nguyên đều được lưu trữ trên các máy chủ, đám mây và người dùng tương tác với những dữ liệu đó thông qua mạng Internet. Chính nhờ sự thuận tiện trong việc giao tiếp, xử lý thông tin đã dẫn đến một hệ lụy nghiệm trọng đó là người dùng hay doanh ngiệp có thể bị đánh cắp dữ liệu từ những website lừa đảo. Các website giả mạo này ngày càng tinh vi và khó thể phát hiện, chỉ cần một cái click chuột cũng có thể khiến cho dữ liệu bị rò rỉ, đây đều là những dữ liệu quan trọng, cần được bảo vệ hàng đầu.

Chúng ta không thể ngăn chặn từng người một truy cập vào các website lừa đảo hay loại bỏ các website này vì số lượng của chúng rất lớn, tuy nhiên ta có thể cảnh báo người dùng trước khi click vào chúng, giúp cho người dùng có thể tự phòng tránh bị mất dữ liệu. Từ đó đề tài mô hình dự đoán website giả mạo được tạo ra để hướng tới các giá trị quan trọng :

– Mục tiêu nhận thức : Việc nghiên cứu này sẽ giúp nhưng người dùng tăng nhận thức hơn về bảo mật an ninh mạng, tránh các rủi ro có thể xảy ra, nâng cao chất lượng không gian mạng.

– Mục tiêu sáng tạo : Đề tài này còn là cơ sở cho các các cá nhân, tổ chức khác phát triển, nâng cao khả năng bảo mật thông tin hay tích hợp vào các phần mềm nhận diện giả mạo tiên tiến hơn...

– Mục tiêu kinh tế : Việc nhận biết được các website lừa đảo sẽ giúp các doanh nghiệp có thể củng cố thông tin của mình, tránh để các dữ liệu đến tay những người lừa đảo vì họ có thể sử dụng thông tin này để làm giàu bản thân và phá hoại nền kinh tế thị trường.

2. Hướng phát triển đề tài

Nghiên cứu này là cơ sở để phát triển mô hình dự đoán lừa đảo được tích hợp vào tiện tích của trình duyệt web (hay còn gọi là extension). Việc đưa vào extension của trình duyệt web giúp cho người dùng có thể dễ dàng cài đặt và sử dụng, đem lại tiện ích và hiệu quả tối đa. Tuy nhiên để tạo extension cũng gặp rất nhiều khó khăn như phải có sự cấp phép của bên trình duyệt hay sự tin tưởng từ phía người dùng. Do đó trong tương lai dự án này cần phải cải thiện nhiều hơn về tính bảo mật, tiện lợi, quảng cáo đến người dùng...

Ngoài phát triển thành extension, ta có thể mô hình hóa nghiên cứu này thành phần mềm được cài đặt trên các thiết bị. Điều này sẽ giúp người dùng có thể dự đoán lừa đảo ở mọi trình duyệt web thay vì chỉ xài được ở một trình duyệt như extension. Tuy vậy thì hướng phát triển này vẫn cần được cải thiện về tính an toàn để có thể được cài đặt trên thiết bị của người dùng mà không xảy ra xung đột với hệ điều hành.

VI. Kết luận

Dựa vào độ chính xác của các thuật toán cũng như ưu và nhược điểm, nhóm đã quyết định sử dụng thuật toán Support Vector Machine (SVM).

Về giá trị khoa học, đề tài hướng đến các giá trị về nhận thức, sáng tạo và kinh tế.

Về hướng phát triển tương lai, dự kiến sẽ phát triển thành extension tích hợp vào trình duyệt nhân chromium, và mô hình hóa thành các phần mềm cài đặt trên các thiết bị di động.

Đây sẽ là một công cụ hỗ trợ đặc biệt cho các người dùng chưa có nhiều kiến thức, kinh nghiệm để phòng tránh. Đặc biệt, do ứng dụng các thuật toán học máy, ta có thể ngày càng nâng cấp mô hình hoàn thiện hơn nhờ vào lượng dữ liệu ngày càng nhiều từ chính bản thân người dùng cũng như cộng động dữ liệu trên thế giới. Dù vậy, đề tài vẫn còn nhiều nhược điểm như sau:

– Các mô hình đưa ra đều có độ chính xác trong khoảng từ 80% đến 90%. Tuy nhiên đây vẫn chưa phải là một con số quá tốt. Việc không tìm được mô hình tốt

hơn, hoặc chưa chuẩn hoá dữ liệu tốt hơn sẽ khiến cho mô hình có sự sai lệch trong dự đoán.

– Chưa có được sản phẩm thực tế để có được trải nghiệm chính xác. Trên thực tế, độ chính xác mô hình cao chưa dẫn đến một sản phẩm tốt, mà còn kết hợp nhiều yếu tố cấu thành.

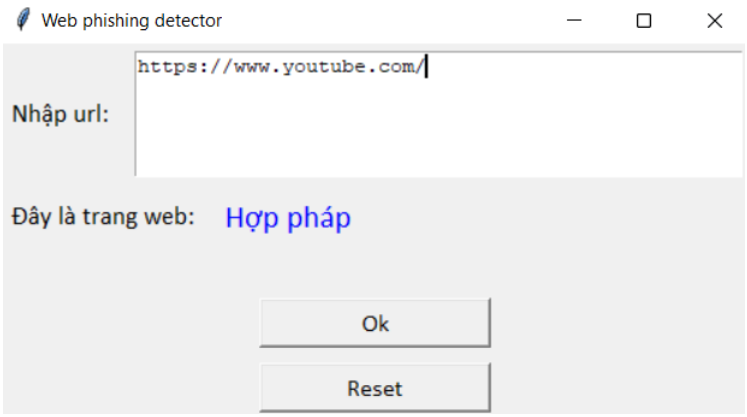
– Đã tồn tại những extension tốt để cảnh báo lừa đảo, chẳng hạn như "Chống lừa đảo" của Hiếu PC.

Trong tương lai, nhóm sẽ cố gắng khắc phục các nhược điểm của đề tài để ngày càng hoàn thiện hơn.

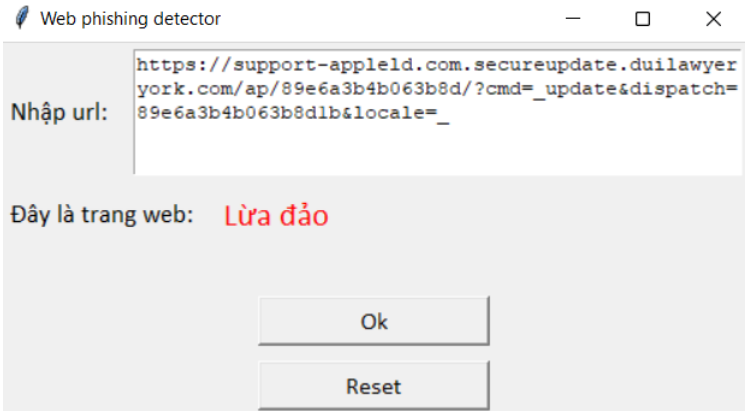
VII. Sản phẩm Demo

Mọi người có thể tham khảo thông qua link: `https://github.com/ToaiTran2001/Web-Phishing-Detector.git`

Nhóm đã làm một desktop app đơn giản để demo mô hình vừa huấn luyện, giao diện của sản phẩm như sau:



Hình 10: Mô hình dự đoán trang web hợp pháp



Hình 11: Mô hình dự đoán trang web lừa đảo

TÀI LIỆU THAM KHẢO

[1] Huynh Chi Trung. (2020). Giới thiệu về Support Vector Machine (SVM). Truy cập từ: https://viblo.asia/p/gioi-thieu-ve-support-vector-machine-svm-6J3ZgPVElmB.

[2] Jian Mao, Jingdong Bian, Wenqian Tian, Shishi Zhu, Tao Wei, Aili Liand Zhenkai Liang. (2019). Phishing page detection via learning classifiers from page layout feature. Truy cập từ: https://jwcn-eurasipjournals.springeropen.com/articles/10.1186/s13638-019-1361-0

[3] Dalia Shihab Ahmed, Assist. Prof. Dr. Karim Q. Hussein, Hanan Abed Alwally Abed Allah. (2022). Turkish Journal of Computer and Mathematics Education. Vol. 13 No. 01. 100 - 107.

[4] Abdelhakim Hannousse, Salima Yahiouche. (2021). Web page phishing detection. Truy cập từ: https://data.mendeley.com/datasets/c2gw7fy2j4/3

[5] Rami M. Mohammad, Fadi Thabtah, Lee McCluskey. (2022). Phishing Websites Features.