The Journal of the Convergence on Culture Technology (JCCT)
Vol. 10, No. 4, pp.575-580, July 31, 2024. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2024.10.4.575
JCCT 2024-7-66

머신러닝 기반 피싱 사이트 탐지 모델

Machine Learning-based Phishing Website Detection Model

오수민*, 박민서**

Sumin Oh*, Minseo Park**

요 약 소셜 미디어의 대중화로 지능화된 피싱 공격을 방어하기 위해 접근하고자 하는 사이트의 상태(정상/피싱)를 판별하는 것이 필요하다. 본 연구에서는 머신러닝 기반 분류 모델을 통해 사이트의 정상/피싱 여부를 예측하는 모델을 제안한다. 첫째, 'URL'에 대한 정보를 수집하여 수치 데이터로 변환한 후, 이상치를 제거한다. 둘째, 변수들 간의 상관관계 및 독립성을 파악하기 위해 VIF(Variance Inflation Factors)를 적용한다. 셋째, 머신러닝 기반 분류 모델을 활용하여 피싱 사이트 탐지 모델을 개발하고, 이를 통해 사이트의 상태를 예측한다. 분류 모델 중 랜덤 포레스트(Random Forest)의 성능이 가장 우수했으며, 테스트 데이터에서 정밀도(Precision) 93.74%, 재현율(Recall) 92.26%, 정확도(Accuracy) 93.14%를 보였다. 향후 이 연구는 다방면의 피싱 범죄 탐지에 적용할 수 있을 것으로 기대된다.

주요어 : 머신러닝, 피싱 사이트, 분류 모델, 랜덤 포레스트

Abstract Detecting the status of websites, normal or phishing, is necessary to defend against intelligent phishing attacks. We propose a machine learning-based classification to predict the status of websites. First, we collect information about 'URL', convert it into numerical data, and remove outliers. Second, we apply VIF(Variance Inflation Factors) to understand the correlation and independence between variables. Finally, we develop a phishing website detection model with machine learning-based classifications, which predicts website status. In the test datasets, Random Forest showed the best performance, with precision of 93.74%, recall of 92.26%, and accuracy of 93.14%. In the future, we expect to apply our model to detect various phishing crimes.

Key words : Machine Learning, Phishing Website, Classification Model, Random Forest

Ⅰ. 서 론

피싱(Phishing)은 개인정보(Private Data)를 낚는다(Fishing)는 의미의 합성어로, 정부, 금융회사 등 신뢰할 수 있는 기관의 웹 사이트(Website)로 위장한 피싱 사이트를 통해 개인정보를 탈취하는 사기 행위이다[1].

피싱 공격은 불특정 다수를 대상으로 이루어지기 때문에 피해를 예측하기 어려우며, 그로 인한 피해의 규모가 광범위하고 피해 정도가 심각하다. 금융감독원에 따르면, 2023년 피싱으로 인한 국내 총 피해액은 1,965억 원으로 전년 대비 35.40% 증가하였다[2].

소셜 미디어(Social Media)의 대중화로 피싱 공격은 더욱 지능적이고 다양해지고 있다[3]. APWG (Anti-Phishing Working Group)에 따르면, 2023년에는 역대 최고 수준인 약 500만 개의 피싱 사이트가 발견되었으며, 그중 24.76%는 소셜 미디어를 기반으로 피싱 공격

*준회원, 서울여자대학교 데이터사이언스학과 학부생
**정회원, 서울여자대학교 데이터사이언스학과 교수
접수일: 2024년 4월 18일, 수정완료일: 2024년 5월 20일
게재확정일: 2024년 6월 10일

Received: Apil 18, 2024 / Revised: May 20, 2024
Accepted: June 10, 2024
**Corresponding Author: mpark@swu.ac.kr
Dept. of Data Science, Seoul Women's Univ, Korea

을 시도하는 것으로 확인되었다[4].

피싱 사이트는 URL(Uniform Resource Locator)을 기반으로 정상 사이트와 구분할 수 있다[5]. 그러나 처음 접근하는 웹 사이트의 URL의 경우, 위장되었는지 판별하는 것이 어려울 수 있다. 이에 이전 접속 여부와 관계없이 URL만을 활용하여 피싱 사이트를 자동으로 판별하는 것이 필요하다.

최근 URL을 활용해 피싱 사이트를 탐지하는 다양한 연구가 활발하게 이루어지고 있다[6-8]. Huang 등[6]은 URL에 포함된 단어를 활용해 피싱 URL을 분류하는 방법을 제안하였다. 단어 임베딩(Word Embedding)을 활용하여 단어 정보를 압축하고, CNN(Convolutional Neural Network) 레이어가 추가된 Attention-based LSTM(Long Short-Term Memory)을 활용하여 피싱 URL에서 다수 추출되는 단어의 패턴을 학습하였다. Ali 등[7]은 URL의 문자(Character) 정보를 기반으로 피싱 URL의 특성을 분석하고 분류하는 방법을 제안하였다. URL을 문자(Character) 단위로 쪼개 각각을 인코딩(Encoding)하여 FCN(Fully Connected Layer)이 추가된 CNN 기반의 분류 모델의 입력으로 활용하였다. Yuan 등[8]은 URL에 포함된 정보를 섹션 별로 구분하여 피싱 URL을 탐지하는 방법을 제안하였다. URL 프로토콜(URL Protocol), 하위 도메인 이름(Sub-domain Name), 도메인 이름(Domain Name), 도메인 접미사(Domain Suffix), URL 경로(URL Path)의 다섯 가지 섹션으로 URL을 구분하고, 각각을 XGBoost(Extreme Gradient Boosting)의 입력으로 활용하였다. 그러나 문자, 단어, 섹션 단위로 패턴을 학습하는 경우, 관측된 적 없는 새로운 패턴을 가진 URL에 대한 탐지 성능이 크게 하락할 수 있다는 한계가 존재한다. 피싱 사이트 탐지의 일반화된 성능을 위해서는 URL에 포함된 다양한 특징을 고려해야 한다.

따라서, 본 연구는 URL의 다양한 특징을 고려해 피싱 사이트를 탐지하는 머신러닝(Machine Learning) 기반의 예측 모델을 제안한다. 머신러닝 기반의 이진 분류 모델을 개발하고, K-겹 교차검증(K-Fold Cross Validation)을 통해 모델의 정확도를 향상시킨다.

본 논문은 다음과 같이 구성된다. 제2장에서는 다양한 머신러닝 기반의 분류 모델에 대해 언급한다. 제3장에서는 모델을 설계하고 검증을 통해 최적의 모델을 탐색한다. 제4장에서는 결과를 분석한다. 마지막으로 제5장에서는 결론과 향후 연구 방향에 대해 언급한다.

II. 머신러닝 기반 분류 모델

피싱 사이트 탐지는 정상 URL과 피싱 URL의 두 카테고리에 대한 예측이므로 머신러닝 기법 중 분류 모델이 적절하다. 본 장에서는 머신러닝 기반의 분류 모델에 대해 살펴보고자 한다.

1. 로지스틱 회귀(Logistic Regression)

로지스틱 회귀(Logistic Regression)는 일반적으로 종속변수가 질병의 유무와 같이 0 또는 1의 값을 가지는 이항 변수일 때 사용하는 모델이다. 회귀계수(Coefficient)를 활용하여 각 독립변수의 중요도를 파악할 수 있다는 장점이 있다. 종속변수가 3개 이상의 결과를 도출하는 경우를 다중 로지스틱 회귀(Multiple Logistic Regression)라고 한다[9].

2. 의사결정 트리 트리(Decision Tree)

의사결정 트리(Decision Tree)는 어떤 조건이 특정 결과를 가져올 것인지 예측하는 의사결정 규칙[10]에 따라 데이터를 분류하는 모델이다. 크고 복잡한 데이터 세트를 효율적으로 처리할 수 있고, 의사결정 과정을 트리의 분기마다 의사결정 규칙이 반영된 트리 모양 구조로 시각화할 수 있다는 장점이 있다[11]. 그러나 하나의 트리 기반 분석을 진행하므로 트리의 깊이가 깊어질 수 있어 과적합(Over-Fitting)이 발생할 수 있다. 이에 적절한 시점에서 가지치기(Pruning)할 필요가 있다[12].

3. 랜덤 포레스트(Random Forest)

랜덤 포레스트(Random Forest)는 많은 데이터에 작은 의사결정 트리(Decision Tree) 여러 개를 동시에 적용하여 조합하는 모델이다. 데이터 내에서 복원 추출을 통해 여러 개의 데이터 세트를 만든 후, 각각의 데이터 세트에 대해 의사결정 트리(Decision Tree)를 생성한다. 각 트리의 결과를 통합시켜 최종 예측을 출력한다[13]. 다양한 데이터 특징을 고려할 수 있어 높은 수준의 예측 정확도 달성이 가능하다. 특히, 변수가 많거나 결측치가 많은 데이터에 좋은 성능을 보인다[14].

III. 연구방법

본 연구는 데이터 수집, 데이터 전처리, 변수 선정, 모델링 과정을 거쳐 머신러닝 기반 피싱 사이트 탐지 모델을 개발한다.

1. 데이터 수집

본 연구에서는 캐글(Kaggle)[15]에서 제공하는 데이터를 사용한다. 본 데이터는 URL, 각 URL에 포함되는 특정 기호의 개수, 도메인 포함 여부, 웹 사이트의 상태(정상/피싱 여부) 등 총 19,431개의 데이터(87개의 변수)로 구성된다.

2. 데이터 전처리

'URL'에 대한 정보 중 중복되는 내용이 포함되는 'URL' 변수를 제거한다. 다음으로 모든 변수가 갖는 값의 형태를 수치 데이터로 통일하고, 이상치(Outliers)를 제거한다. '저작권 문구가 있는 도메인 여부'와 '페이지 상태'의 경우, 데이터의 형태가 문자형이기 때문에 이진화(Binarization)를 통해 수치 데이터로 변환한다. 이상치를 포함하는 변수인 '도메인 연령'은 도메인 혹은 URL이 인터넷에 등장한 지 얼마나 오래되었는지를 나타내는 변수이다. 도메인이 생성된 직후부터 0 이상의 값을 가지므로, 0 미만의 값을 갖는 행은 이상치로 판단하여 제거한다. 이상치를 제거하여 남은 행은 총 14,137개이다.

3. 변수 선정

전처리를 통해 정제된 데이터 안에서 종속변수인 페이지 상태를 제외한 84개 변수를 사용한다. 변수들의 독립성에 대한 유의성을 검정하기 위해 VIF(Variance Inflation Factor, 분산팽창지수)를 살펴본다. VIF는 해당 변수가 다른 변수와 어떤 상관관계를 가지는지, 독립성을 띠는지를 정량화한 값이다[16]. 일반적으로 10이 넘으면 독립변수 간 상관관계가 존재하며, 하나의 독립적인 변수 역할을 하기 어렵다고 판단한다. 이에 VIF 10 이상의 값을 갖는 '호스트 네임의 평균 단어 길이', 'URL의 단어 수', '호스트 네임의 가장 긴 단어의 길이' 등 25개의 독립변수를 제거한다. 변수 선정 과정을 통해 모델 학습에 사용될 59개의 독립변수를 도출한다. 각 독립변수 간 VIF는 표 1과 같다.

표 1. 독립변수의 분산팽창지수
Table 1. Variance Inflation Factors of Independent Variables

독립변수	VIF	독립변수	VIF
Page Rank	8.1	Shortening Service	1.6
Domain Age	7.8	Web Traffic	1.5
Domain In Title	4.8	Abnormal Subdomain	1.4
Ratio ExtHyperlinks	4.7	Ratio ExtErrors	1.4
Links In Tags	4.5	Total Of '_'	1.3
Google Index	3.5	Nb ExtCSS	1.3
Ratio IntMedia	3.3	Total Of '@'	1.3
External Favicon	3.2	Total Of Http In Path	1.2
Https Token	3.2	Login Form	1.1
Total Of Com	3.2	Total Of ';'	1.1
Ip	3.0	Random Domain	1.1
Safe Anchor	2.8	Dns Record	1.1
Total of 'www'	2.5	Total Of '%'	1.1
Tld In Subdomain	2.5	Brand In Path	1.1
Ratio Digits Url	2.5	Statistical Report	1.1
Ratio ExtMedia	2.4	Suspecious Tld	1.1
Domain With Copyright	2.2	Total Of '~'	1.1
Tld In Path	2.1	Whois Registered Domain	1.1
Domain In Brand	2.1	Brand In Subdomain	1.1
Empty Title	2.1	Onmouseover	1.0
Total Of '?'	2.0	Total Of '$'	1.0
Nb Redirection	1.9	Total Of ','	1.0
Ratio Digits Host	1.9	Total Of '*'	1.0
Shortest Word Path	1.9	Popup Window	1.0
Char Repeat	1.8	Port	1.0
Nb Hyperlinks	1.8	Right Clic	1.0
Domain Registration Length	1.7	Iframe	1.0
Ratio ExtRedirection	1.7	Punycode	1.0
Prefix Suffix	1.6	Path Extension	1.0
Phish Hints	1.6		

4. 모델링

데이터 전처리와 변수 선정 과정을 거친 데이터를 로지스틱 회귀(Logistic Regression), 의사결정 트리(Decision Tree), 랜덤 포레스트(Random Forest)의 세 가지 분류 모델에 적용한다. 실험을 위하여 훈련 데이터(Training Sets)와 테스트 데이터(Test Sets)를 무작위로 각각 80%, 20%의 비율로 나누어 구성한다.

의사결정 트리(Decision Tree)의 노드 분할 시, 지니 불순도(Gini Impurity)를 각 노드의 분류 기준으로 사용한다[17]. 과적합(Over-fitting)을 방지하기 위해[18] 트리의 최대 깊이는 3으로 제한한다. 랜덤 포레스트(Random Forest)를 구성하는 결정 트리의 개수는 100개, 노드 분할 기준은 엔트로피(Entropy), 트리의 깊이를 3으로 제한한다. 탐지 모델의 정확도를 검증하기 위하여 10-겹 교차 검증(10-Fold Cross Validation)을 수행한다. 그림 1은 제안 모델의 의사코드(Pseudo Code)이다.

```
1. Split Data ; X ( Various features of URL ), y ( phishing, legitimate )
        train_test_split(X, y, test_size=0.2)
2. Modeling ; Train three different models
        Logistic Regression
        Decision Tree ( criterion='gini', max_depth=3 )
        Random Forest ( n_estimators=100, criterion='entropy', max_depth=3 )
3. Evaluate Model Performance
        model_evaluate = accuracy_score(y_test, model_test_pred)
4. 10-Fold Cross Validation
        kf = KFold(n_splits=10, shuffle=True)
5. Print Model Performance ; Precision, Recall, Accuracy, F1-Score
```

그림 1. 피싱 사이트 탐지 모델의 의사코드
Figure 1. Pseudo code of phishing site detection model

Ⅳ. 실험 및 결과 분석

탐지 모델의 설명력을 정량적으로 평가하고 검증하기 위해 모델의 예측력을 나타내는 정밀도(Precision), 재현율(Recall), 정확도(Accuracy), F1-Score를 측정하였다. 표 2는 피싱 사이트를 탐지하는 모델의 정량적 평가 결과이다.

표 2. 피싱 사이트 탐지 모델의 정량적 평가
Table 2. Quantitative evaluation of phishing site detection models

Model	Datasets	Evaluation (%)			
		Precision	Recall	Accuracy	F1-Score
Logistic Regression	Training	70.98	50.39	65.50	58.94
	10-fold Cross Validation	70.86	92.91	65.47	59.09
	Test	70.43	51.68	65.42	59.62
Decision Tree	Training	91.98	89.83	91.16	90.89
	10-fold Cross Validation	92.12	92.91	91.15	90.88
	Test	92.56	89.11	91.09	90.80
Random Forest	Training	**93.24**	**93.10**	**93.30**	**93.28**
	10-fold Cross Validation	**93.35**	**92.91**	**93.24**	**92.64**
	Test	**93.74**	**92.26**	**93.14**	**93.25**

피싱 사이트 탐지 모델의 모든 평가 수치에서 각각의 훈련 및 테스트 데이터, 10-겹 교차 검증(10-Fold Cross Validation)의 성능 차이가 크게 없으므로 세 모델 모두 일반화된 성능을 가짐을 확인하였다. 테스트 데이터에서 세 가지 분류 모델 중 랜덤 포레스트(Random Forest) 모델이 정밀도(Precision) 93.74%, 재현율(Recall) 92.26%, 정확도(Accuracy) 93.14%, F1-Score 93.25%로 가장 우수한 성능을 보이며, 이를 통해 랜덤 포레스트(Random Forest) 모델이 상대적으로 정확하게 피싱 사이트를 탐지하는 것을 알 수 있었다.

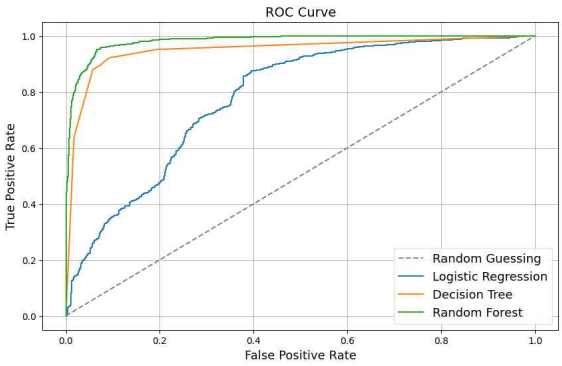

그림 2. 피싱 사이트 탐지 모델 각각의 ROC Curve
Figure 2. Receiver operating characteristic curve of each phishing site detection model

그림 2는 테스트 데이터를 기반으로 나타낸 세 가지 분류 모델의 ROC Curve(Receiver Operating Characteristic Curve)이다. ROC Curve는 분류 모델의 성능을 시각적으로 나타낸 그래프로, 커브의 아래 면적은 정확도를 나타낸다. 이때, 곡선이 좌측 상단에 가까울수록 분류 모델의 성능이 좋은 것으로 판단한다[19]. 파란색 선은 로지스틱 회귀(Logistic Regression), 주황색 선은 의사결정 트리(Decision Tree), 초록색 선은 랜덤 포레스트(Random Forest)의 ROC Curve이다. ROC Curve를 이용해 피싱 사이트 탐지 모델의 정량적 평가 지표를 시각화한 결과, 초록색 선인 랜덤 포레스트(Random Forest)의 성능이 가장 우수한 것을 확인할 수 있었다.

Ⅴ. 결 론

본 논문에서는 머신러닝(Machine Learning) 기반의 피싱 사이트(Phishing Website) 탐지 모델을 제안하였다. URL(Uniform Resource Locator)과 URL 관련 정보를 담고 있는 데이터, 웹 사이트(Website)의 상태(정상/피싱)를 수집하였다. 문자형 변수를 수치형으로 변환하고, 모델의 예측력에 문제가 될 수 있는 이상치(Outliers)를 제거하였다. 유의한 독립변수를 탐색하기 위해 VIF(Variance Inflation Factor, 분산팽창지수)를 기준으로 유의성 평가를 진행하여 변수를 다시 한번 전처리하였다. 선정된 변수 59개를 머신러닝의 분류 모델인 로지스틱 회귀(Logistic Regression), 의사결정 트리(Decision Tree), 랜덤 포스트(Random Forest)에 적용하고, 10-겹 교차검증(10-Fold Cross Validation)을 수행하였다. 실험 결과, 복원 추출을 통해 다양한 데이터

특징을 반영하여 예측을 출력하는 모델인 랜덤 포레스트(Random Forest) 모델이 가장 우수한 성능을 가짐을 확인하였다.

제안 방법은 사이트의 URL 관련 정보를 바탕으로 정상/피싱 사이트를 구분하는 모델로, 금융기관 홈페이지 등을 가장한 피싱 사이트의 URL 주소 차이를 육안으로 확인할 필요 없이 자동으로 분류할 수 있어 SNS를 통해 연결된 피싱 사이트의 접속을 사전에 차단할 수 있다는 의의가 있다. 향후 연구로는 URL을 이용한 피싱뿐 아니라 다른 종류의 사이버 피싱 데이터를 추가적으로 수집하여, 다방면의 피싱 범죄 탐지에 적용해 보고자 한다.

References

[1] N. Suryavanshi and A. Jain, "A Review of Various Techniques for Detection and Prevention for Phishing Attack," *International Journal of Advanced Computer Technology (IJACT)*, Vol. 4, No. 3, pp. 41-46, 2015.

[2] Financial Supervisory Service, 2024. Available online: www.fss.or.kr (accessed on 04 April 2024)

[3] J. Yoon and S. Buu, "Deep Character-level Anomaly Detection based on a Transformer-style Convolutional Autoencoder for Phishing URL Detection," *Proceedings of KIIT Conference*, pp. 114-118, 2023.

[4] APWG, 2024. Available online: https://apwg.org/ (accessed on 04 April 2024)

[5] Police Department, 2024. Available online: www.police.go.kr (accessed on 04 April 2024)

[6] Y. Huang, Q. Yang, J. Qin, and W. Wen, "Phishing URL Detection via CNN and Attention-Based Hierarchical RNN," *IEEE International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications*, pp. 112-119, 2019.

[7] A. Ali, Q. Jiang, Q. Qu, M. Huang, and J. P. Niyigena, "An effective phishing detection model based on character level convolutional neural network from URL," *Electronics*, Vol. 9, No. 9, 2020. DOI: 10.3390/electronics9091514

[8] H. Yuan, Z. Yang, X. Chen, Y. Li, and W. Liu, "URL2Vec: URL modeling with character embeddings for fast and accurate phishing website detection," *IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications*, pp. 265-272, 2018.

[9] D. G. Kleinbaum and M. Klein, "Logistic Regression: A Self-Learning Text," 2010. DOI: 10.1007/978-1-4419-1742-3

[10]S. Greco, B. Matarazzo, and R. Słowiński, "Decision Rule Approach," *Multiple criteria decision analysis: state of the art surveys*, pp. 497-552, 2016.

[11]D. J. Hand, "Principles of Data Mining," *Drug safety*, Vol. 30, pp. 621-622, 2007.

[12]S. J. Rigatti, "Random Forest," *Journal of Insurance Medicine*, Vol. 47, No. 1, pp. 31-39, 2017. DOI: 10.17849/insm-47-01-31-39.1

[13]Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry*, Vol. 27, No. 2, pp. 130-135, April 2015. DOI: 10.11919/j.issn.1002-0829.215044

[14]L. Breiman, "Random Forest," *Machine learning*, Vol. 45, pp. 5-32, October, 2001. DOI: 10.1023/A:1010933404324

[15]Kaggle, 2024. Available online: https://www.kaggle.com/ (accessed on 04 April 2024)

[16]J. K. Harris, "Primer on binary logistic regression," *Family medicine and community health*, Vol. 9, Suppl. 1, 2021. DOI: 10.1136/fmch-2021-001290

[17]J. L. Grabmeier and L. A. Lambe, "Decision trees for binary classification variables grow equally with the Gini impurity measure and Pearson's chi-square test," *International journal of business intelligence and data mining*, Vol. 2, No. 2, pp. 213-226, June 2007. DOI: 10.1504/IJBIDM.2007.013938

[18]S. Sohn, H. Yang, and M. Park, "Analysis of Risk Factors for Youth Population Outflow in Busan Based on Machine Learning," *The Journal of the Convergence on Culture Technology (JCCT)*, Vol. 9, No. 6, pp. 131-136, November 2023, DOI:10.17703/JCCT.2023.9.6.131

[19]Z. H. Hoo, J. Candlish, and D. Teare, "What is a ROC curve?," *Emergency Medicine Journal*, Vol. 34, No. 6, pp. 357-359, May 2017.

※ 이 논문은 서울여자대학교 학술연구비의 지원에 의한 것임 (2024-0112).