웹사이트 피싱 분석

덕성여자대학교

수학과

백재이 (외 4명)

요약
1. 서론
2. 탐색적 데이터 분석
 2.1 범주형 변수
 2.2 연속형 변수
 2.3 연관성 분석
3. 머신러닝 모델링
 3.1 분류 모형 평가 방법
 3.2 로지스틱 회귀
 3.3 서포트벡터 머신
 3.4 의사결정나무
 3.5 랜덤포레스트
 3.6 인공신경망
4. 결과
5. 결론
참고 문헌
부록 : R 코드

요약

최근 피싱기법이 실제와 구분하기 어려울 정도로 정교해지고, 이에 대한 우려가 커지고 있다. 과학기술정보통신부와 한국인터넷진흥원에서도 피싱에 대한 위험성을 인지하여 피싱 예방 대책을 발표하였다. 피싱 예방의 중요성이 증가함에 따라, 본 논문에서는 기계학습(Machine Learning)의 한 방법인 지도학습을 통해서 웹사이트의 피싱 여부를 판단할 수 있는 모델을 제시하였다. 분석 단계에서는 다섯 가지의 지도학습 방법으로 모형을 학습하였고, 모형들을 분류 평가 기준을 따라 비교하였다. 최종적으로 랜덤포레스트를 통해 만들어진 모형이 선택되었고, 어떠한 요인(변수)이 웹사이트의 피싱 여부에 영향을 미치는지 알아보고자 한다.

1. 서론

미국 연방수사국 FBI 산하에 있는 IC3(Internet Crime Complaint Center)의 2021년 보고서에 따르면, 코로나19로 인해 원격 근무와 인터넷 기반 소통이 활발해지면서 가상 세계 의존성을 활용한 사이버 공격이 증가하였다. (IC3, 2021) 그리고 과학기술정보통신부와 한국인터넷진흥원은 국내 주요 보안기업과 함께 코로나19가 지속되고 디지털 전환이 가속화됨에 따라 지능화 및 고도화되는 사이버 위협으로부터 선제적인 예방과 대비태세를 강화하기 위해 2022년 사이버 위협 전망을 발표하기도 했다. (과학기술정보통신부 & 한국인터넷진흥원, 2022) 이처럼 지능화되는 사이버 공격을 막기 위한 정부 차원의 대비와 정교한 예방이 필요성이 커지고 있다.

따라서 이번 분석에서는 피싱 여부를 예측하기 위해 웹사이트에 관한 데이터를 로지스틱 회귀, 서포트벡터 머신, 의사결정나무, 랜덤포레스트, 인공신경망을 이용하여 모델링하여, 그중 가장 뛰어난 성능을 가진 모형을 최종 모형으로 선정하였다.

이후의 논문 구성은 다음과 같다. 2장에서는 웹사이트 피싱 분석 데이터에 대한 설명과 범주형 변수의 빈도, 연속형 변수의 기초통계량, 그리고 종속변수와 독립변수 간의 연관성을 분석하였다. 3장에서는 모델을 평가할 기준에 대한 설명과 각각의 기법으로 모델링하고, 최적의 모형을 선택한다. 4장에서는 최종으로 선택한 모형에 대하여 설명한다. 마지막으로 5장에서는 결론을 제시하고자 한다.

2. 탐색적 데이터 분석

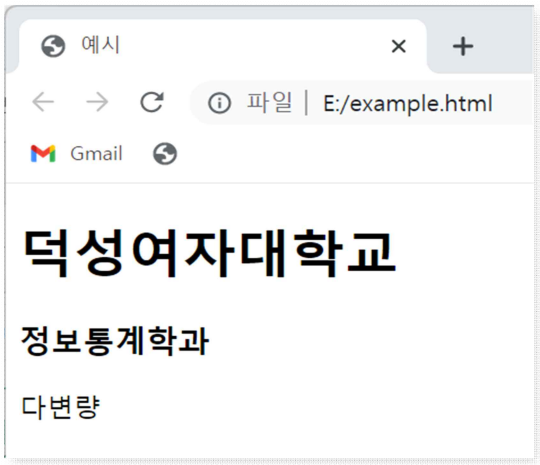
본 분석에 활용한 피싱 데이터는 웹사이트의 정보가 담긴 URL과 웹사이트 기본 구조를 세우는 HTML 소스 코드의 특징으로 구성되어 있다. 논문에 관한 이해를 돕기 위해 해당 분야에 대한 간단한 설명을 덧붙이고자 한다.

웹 통신 구조는 클라이언트와 서버가 응답을 주고받는 구조로 되어 있다. 각자 인터넷이 가능한 환경에서 스마트폰이나 PC 화면을 통해 볼 수 있는 다양한 웹사이트는 서버와의 수많은 응답의 결과물이다. 우리가 웹사이트를 공유할 때 사용하는 URL은 클라이언트와 서버가 인터넷상에서 데이터를 주고받기 위한 약속인 프로토콜, 원하는 자원이 있는 서버의 주소, 그리고 자원의 위치 정보를 담고 있어 클라이언트가 보고자 하는 웹사이트에 접근할 수 있게 한다. 예를 들어, 위키피디아의 피싱에 관한 URL “https://en.wikipedia.org/wiki/Phishing”은 https라는 프로토콜, 위키피디아 영어 버전 도메인 주소, 해당 주소의 wiki 폴더 내 피싱

(Phishing) 파일의 위치가 순서대로 담겨있음을 확인할 수 있다. 그렇게 URL을 통해 접근한 웹사이트는 웹사이트 뼈대 틀을 담당하는 HTML, 디자인 스타일을 담당하는 CSS, 사용자와 반응하는 동적 처리를 구현하는 JavaScript(자바스크립트) 프로그래밍이 조합된 결과물이다.

웹사이트 구조를 담당하는 HTML의 간단한 예시는 <그림 2.1>과 같다. 부등호 모양의 괄호(<>)로 열고 닫는 태그가 트리 구조로 중첩되어 웹사이트에 제목 본문, 하이퍼링크, 이미지 등의 요소가 채워진다.

```
<!DOCTYPE html>
<html lang="ko">
<head>
        <title>예시</title>
</head>
<body>
        <h1>덕성여자대학교</h1>
        <h3>정보통계학과</h3>
        <p>다변량</p>
</body>
</html>
```



<그림 2.1> HTML 예시

본 분석에서 사용한 데이터는 2015년 1월부터 5월까지, 2017년 5월부터 6월까지 다운로드된 5000개의 피싱 웹사이트와 5000개의 합법 웹사이트에서 추출한 48개의 기능을 포함한 것으로, 출처는 Kaggle이다.

분석에 필요한 변수를 선택하기 위해서, 통계적 유효성뿐만 아니라 해당 분야 전문가 지식에 따른 중요도가 같이 고려되어야 한다. 웹 개발 프로세스나 관련 언어인 HTML에 관한 지식이 부족하므로 해당 피싱 데이터를 사용하여 Hybrid Ensemble Feature Selection(HEFS)라는 새로운 변수 선택 방법을 제시하고 있는 논문(Kang L. C.외 4인, 2019)을 바탕으로 10개의 변수를 선택하여 전체 48개의 변수와 비교해보았다. 전체 변수의 중요도 그래프를 보았을 때, 논문을 바탕으로 선택한 11개의 변수가 대부분 포함되어 효율성을 위해 1개의 범주형 종속변수, 5개의 범주형 독립변수, 5개의 연속형 독립변수를 선택하였다. 종속변수인 CLASS_LABEL은 피싱에 대한 변수로 이름을 Phishing으로 변경하여 진행하였다.

<표 2.1> 변수 설명

종류	영문명	변수명	설명
종속변수	Phishing	피싱	0 = 합법 1 = 피싱
범주형	SubmitInfoToEmail	메일 보내기 기능	0 = 미포함 1 = 포함
	FrequentDomainNameMismatch	도메인 이름 불일치	0 = 일치 1 = 불일치
	PctExtResourceUrlsRT	외부 리소스 코드 비율	-1 : < 22% 0 : 22% ~ 61% 1 : > 61%
	ExtMetaScriptLinkRT	외부 파일 코드 비율	-1 ; < 17% 0 : 17% ~ 81% 1 : > 81%
	PctExtNullSelfRedirectHyperlinksRT	비정상 동작 처리 하이퍼링크 코드 비율	-1 : < 31% 0 : 31% ~ 67% 1 : > 67%
연속형	NumDash	대시 기호 수	대시 기호 : -
	NumNumericChars	문자형 숫자 수	
	PctExtHyperlinks	하이퍼링크 코드 비율	HTML 소스 코드에서 하이퍼링크의 백분율
	NumSensitiveWords	민감한 단어 수	민감한 단어 : secure, account, webscr, login, ebaysapi, signin, banking, confirm
	PctNullSelfRedirectHyperlinks	비정상 링크 하이퍼링크 코드 비율	HTML 소스 코드에 빈값, "#"과 같은 자체 리디렉션 값, 현재 웹사이트의 URL 또는 "file://E:/"과 같은 일부 비정상 값이 포함된 하이퍼링크 필드의 백분율

범주형 변수 중 레이블을 통일한 변수는 외부 리소스 코드 비율, 외부 파일 코드 비율, 비정상 동작 처리 하이퍼링크 코드 비율이다. 외부 리소스 코드 비율은 HTML 소스 코드에 외부 이미지, 영상 등을 가져오는 코드의 비율 규칙을 적용하였고, 외부 파일 코드 비율은 HTML 소스 코드에 외부 문서를 가져오는 스크립트 및 링크 태그와 기본 정보를 제공하는 메타 태그의 비율 규칙을 적용하였고, 비정상 동작 처리 하이퍼링크 코드 비율은 HTML 소스 코드에 서로 다른 도메인 이름을 사용하거나 "#"으로 시작하거나 "JavaScript ::void(0)"를 사용하는 하이퍼링크 비율 규칙을 적용하였다. (Mohammad R. M. A, McCluskey L. & Thabtah F., 2015)

2.1 범주형 변수

<표 2.1.1>을 보면, 메일 보내기 기능을 포함하지 않는 경우와 도메인 이름이 일치하는 경우가 그렇지 않은 경우보다 많다. 외부 리소스 코드 비율이 61%를 넘는 경우, 비정상 동작 처리 하이퍼링크 코드 비율이 31% 미만인 경우가 가장 많고, 외부 파일 코드 비율은 비율에 따른 빈도 차이가 비교적 적다.

<표 2.1.1> 범주형 자료들의 빈도

	메일 보내기 기능	도메인 이름 불일치
0	8712	7847
1	1288	2153

	외부 리소스 코드 비율	외부 파일 코드 비율	비정상 동작 처리 하이퍼링크 코드 비율
-1	2808	3988	6094
0	851	2139	953
1	6341	2873	2953

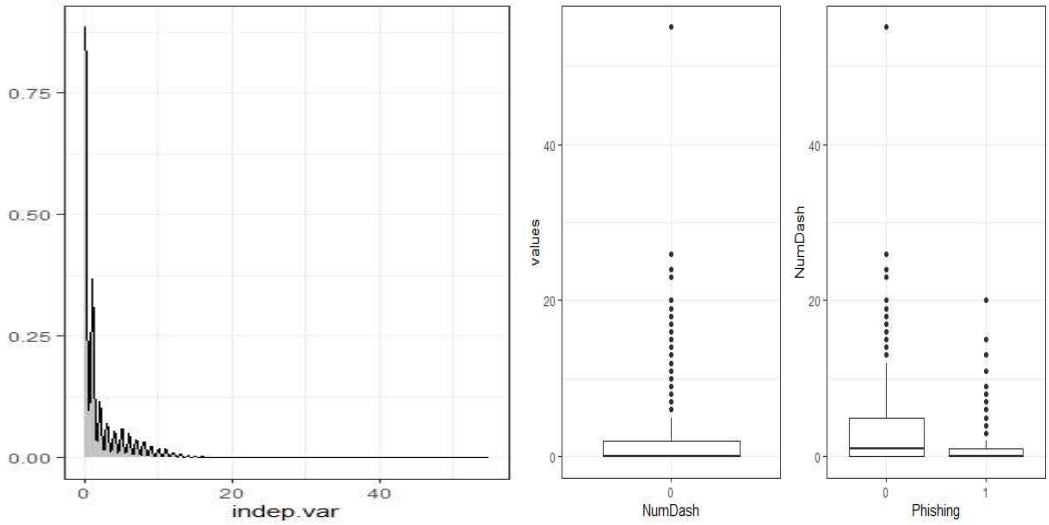
2.2 연속형 변수

<표 2.2.1>은 연속형 변수의 기초통계량으로 각 변수의 관측 개수, 최솟값, 제 1분위수, 중앙값, 평균, 제 3분위수, 최댓값이다. 연속형 변수들의 단위가 다르므로 평균이 0, 표준편차가 1이 되게 표준화하여 분석을 진행했다.

<표 2.2.1> 연속형 변수 기초통계량

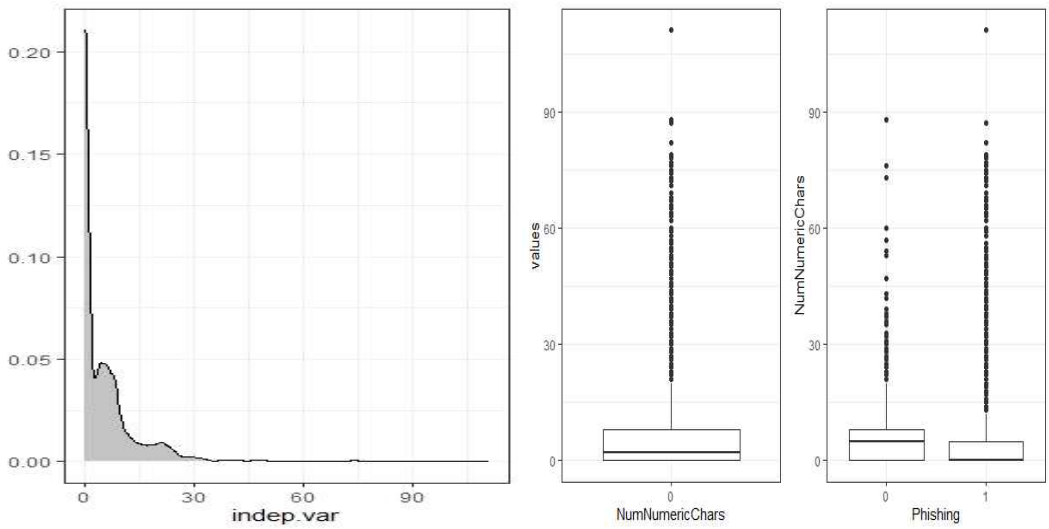
	대시 기호 수	문자형 숫자 수	민감한 단어 수	하이퍼링크 코드 비율	비정상링크 하이퍼링크 코드 비율
관측 개수	10,000	10,000	10,000	10,000	10,000
최솟값	0	0	0	0	0
제 1분위수	0	0	0	0	0
중앙값	0	2	0	0.07	0
평균	1.82	5.81	0.11	0.24	0.14
제 3분위수	2	8	0	0.32	0.05
최댓값	55	111	3	1	1

첫 번째 연속형 변수는 대시 기호 수이다. <그림 2.2.1>의 왼쪽 확률밀도함수 그래프를 보면 웹사이트 내 대시 기호의 수는 대부분 0에서 20 사이의 값을 가지고 있다. 오른쪽 상자 그림에서는 피싱 웹사이트일수록 데이터 분포의 범위가 작음을 알 수 있다. 따라서 피싱 웹사이트일수록 대시 기호의 수가 합법 웹사이트인 경우보다 적다고 할 수 있다.



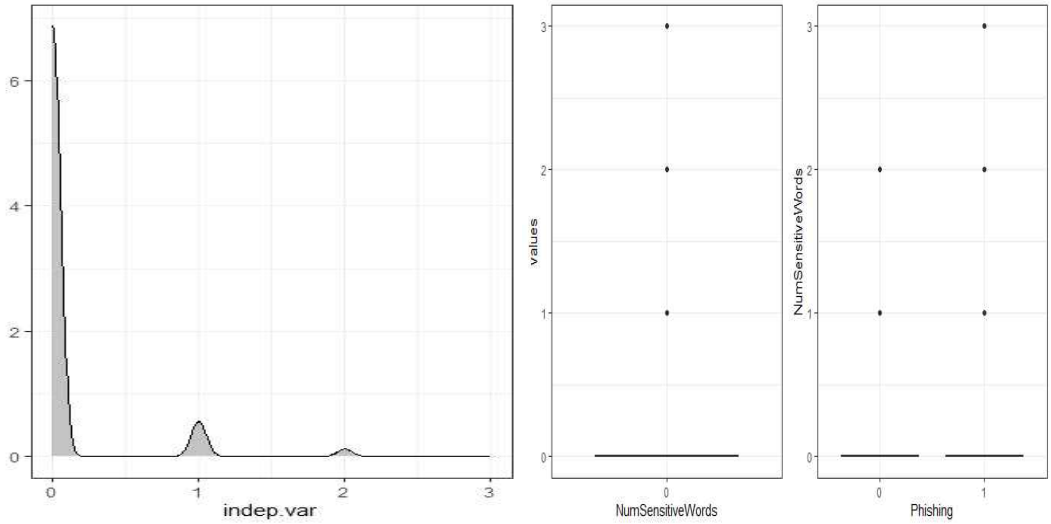
<그림 2.2.1> 대시 기호 수의 분포

두 번째 연속형 변수는 문자형 숫자 수이다. <그림 2.2.2>의 왼쪽 확률밀도함수 그래프를 보면 웹사이트 내의 문자형 숫자 수가 대부분 0개에서 30개 사이인 것을 확인할 수 있다. 오른쪽 상자 그림을 통해 합법 웹사이트인 경우와 피싱 웹사이트인 경우 모두 문자형 숫자 수의 이상치가 많음을 알 수 있다. 특히 피싱 웹사이트인 경우가 합법 웹사이트인 경우보다 많음을 알 수 있다.



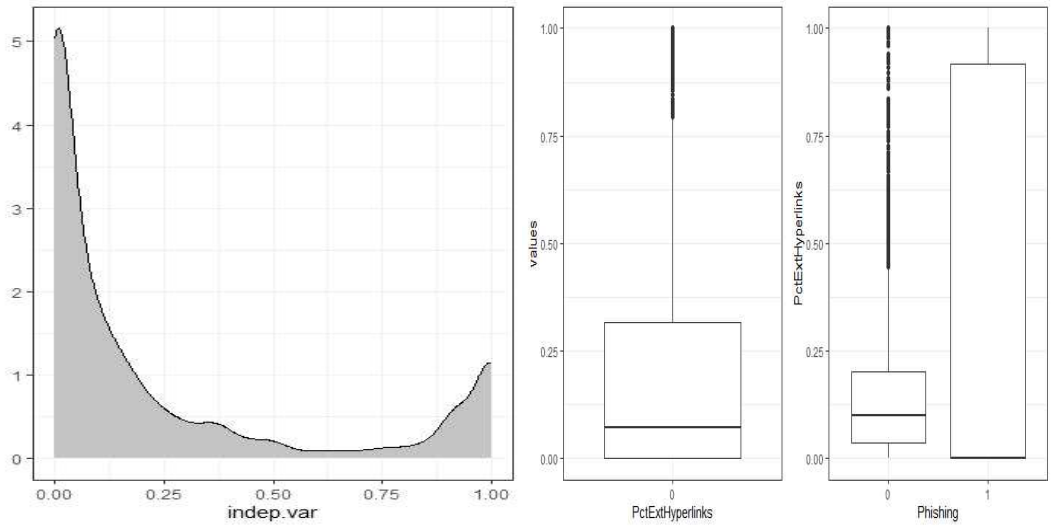
<그림 2.2.2> 문자형 숫자 수의 분포

세 번째 연속형 변수는 민감한 단어 수이다. <그림 2.2.3>의 왼쪽 확률밀도함수 그래프를 보면 민감한 단어의 수는 0개부터 2개까지만 있다는 것을 확인할 수 있다. 오른쪽 상자 그림을 보면 합법 웹사이트인 경우와 피싱 웹사이트인 경우의 분포는 큰 차이를 보이지 않는다.



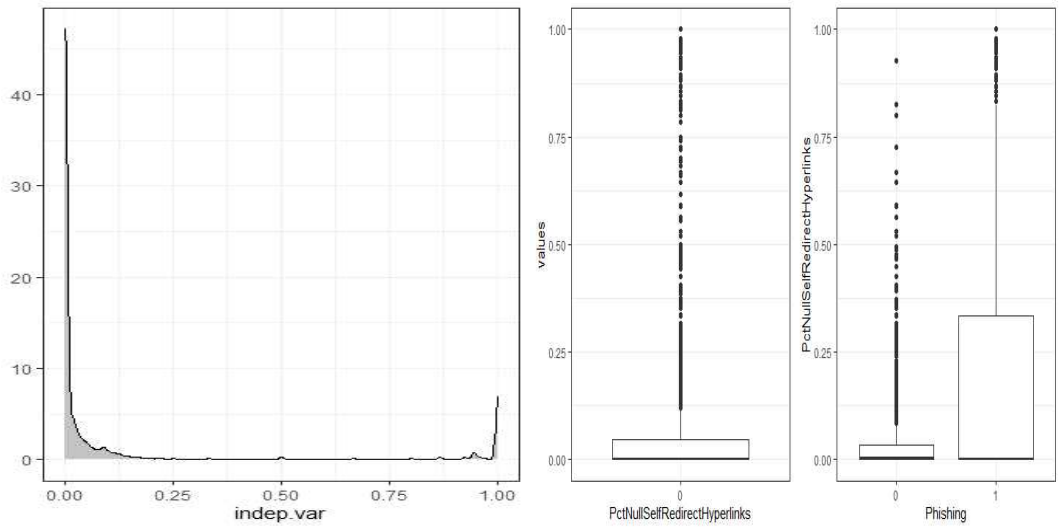
<그림 2.2.3> 민감한 단어 수의 분포

네 번째 연속형 변수는 하이퍼링크 코드 비율이다. <그림 2.2.4>는 하이퍼링크 코드 비율의 확률밀도함수 그래프와 상자 그림이다. 상자 그림을 보면 합법 웹사이트인 경우보다 피싱 웹사이트인 경우의 분포가 더 넓지만, 합법 웹사이트인 경우에는 이상치가 많은 것을 볼 수 있다.



<그림 2.2.4> 하이퍼링크 코드 비율의 분포

다섯 번째 변수는 비정상링크 하이퍼링크 코드 비율이다. <표 2.2.1>를 보면, 비정상링크 하이퍼링크 코드 비율의 중앙값이 0, 평균이 0.14이다. <그림 2.2.5>의 왼쪽 확률밀도함수 그래프로 보아 비율의 대부분이 0임을 알 수 있다. 또한 <그림 2.2.5>의 오른쪽 상자 그림을 보면 이상치가 많이 있고, 합법 웹사이트인 경우보다 피싱 웹사이트인 경우의 분포가 더 넓음을 알 수 있다.

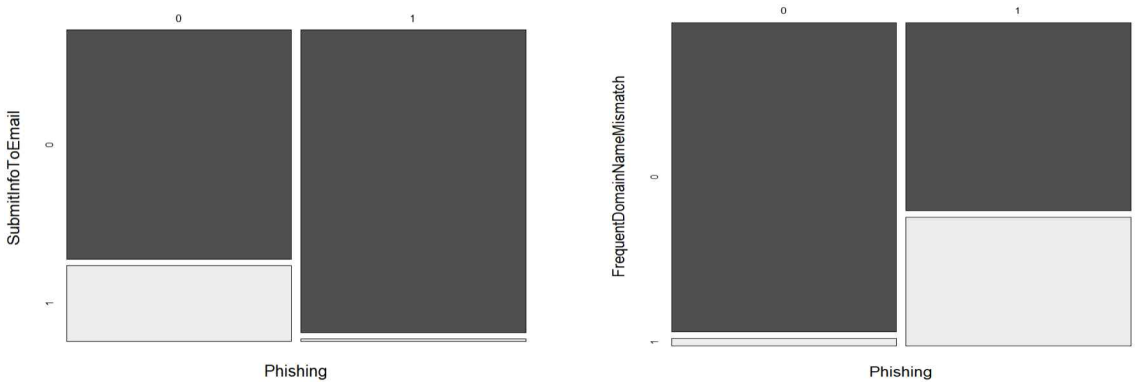


<그림 2.2.5> 비정상링크 하이퍼링크 코드 비율의 분포

2.3 연관성 분석

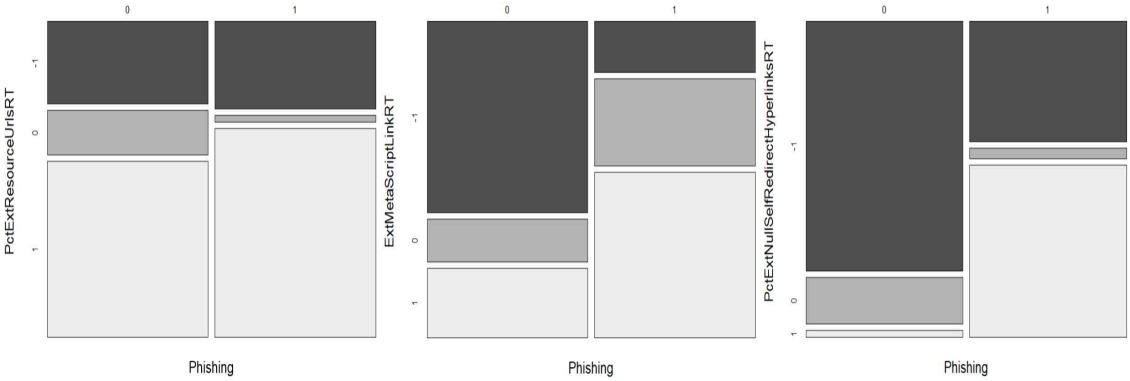
2.3.1 범주형 변수

다음은 범주형 독립변수와 종속변수의 연관성을 알아보기 위해 빈도표를 시각화한 모자이크 그림이다. <그림 2.3.1>은 메일 보내기 기능, 도메인 이름 불일치와 피싱 여부의 모자이크 그림으로, 두 변수의 범주에 따라 피싱 여부에 차이가 있음을 확인할 수 있다.



<그림 2.3.1> 메일 보내기 기능, 도메인 이름 불일치

<그림 2.3.2>는 외부 리소스 코드 비율, 외부 파일 코드 비율, 비정상 동작 처리 하이퍼링크 코드 비율과 피싱 여부의 모자이크 그림으로, 다른 두 변수와 다르게 외부 리소스 코드 비율은 피싱 여부에 따른 차이가 작아 보인다.



<그림 2.3.2> 외부 리소스 코드 비율, 외부 파일 코드 비율, 비정상 동작 처리 하이퍼 링크 코드 비율

카이제곱 검정(Chi-square test) 또는 피셔 정확 검정(Fisher's exact test)으로 각 범주에 따른 피싱 여부의 통계적 유의성을 확인할 수 있다. 본 분석에 활용한 데이터는 각 범주에 따른 빈도표에 기대빈도가 5 이하인 칸이 없으므로, 관찰된 빈도와 기대되는 빈도가 유의미하게 차이가 있는지 검정하기 위해 카이제곱 분포에 기초한 통계적 방법인 카이제곱 검정을 하며, 연구 가설은 다음과 같다.

H_0 : 각 범주형 변수와 피싱 여부 사이에 연관성이 없다.
H_1 : 각 범주형 변수와 피싱 여부 사이에 연관성이 있다.

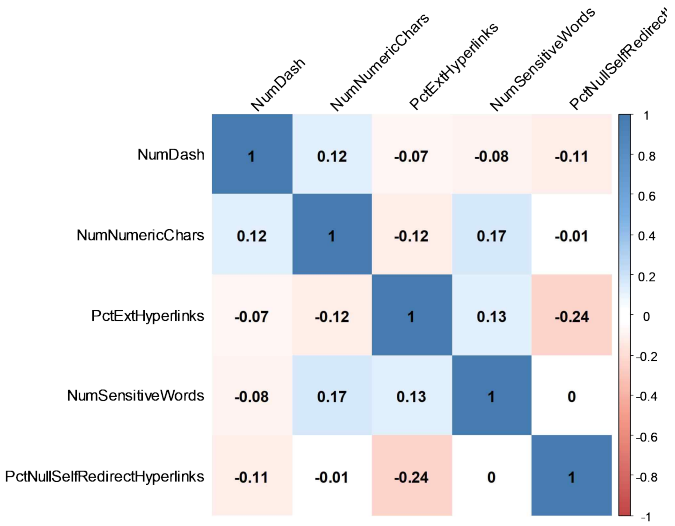
<표 2.3.1>은 각 범주형 독립변수와 종속변수 간의 빈도표와 카이제곱 검정 결과를 정리한 표이다. 모든 변수에 대해서 p-값이 0.05보다 작으므로, 유의수준 0.05 하에서 각 범주형 독립변수와 종속변수인 피싱 여부 사이에 연관성이 있다고 판단한다. 따라서 모든 범주형 변수를 분석에 포함한다. HTML 소스 코드에서 가장 자주 사용하는 도메인 이름과 웹사이트 URL 도메인 이름이 일치하지 않는 경우, 비정상 동작 처리 하이퍼링크 코드 비율이 67%를 초과하는 경우에 피싱인 비율이 0.9 이상으로 피싱 웹사이트의 대표적인 특징임을 알 수 있다. 반대로 메일 보내기 기능이 있는 경우에 피싱인 비율은 0.03으로 제일 작으며, 웹사이트 호스트가 클라이언트로부터 메일을 받기 위해 메일 보내기 기능 코드에 공개하는 메일 주소는 오히려 피싱 또는 스팸 공격의 대상이므로 피싱 웹사이트에 잘 쓰이지 않는다.

<표 2.3.1> 범주형 변수 카이제곱 분석

		피싱		전체	χ^2	p
		합법	피싱			
메일 보내기 기능	미포함	3757	4955	8712	1279.028	<0.0001
	포함	1243	45	1288		
도메인 이름 불일치	일치	4877	2970	7847	2152.552	0
	불일치	123	2030	2153		
외부 리소스 코드 비율	< 22%	1363	1445	2808	511.198	<0.0001
	22% ~ 61%	739	112	851		
	> 61%	2898	3443	6341		
외부 파일 코드 비율	< 17%	3151	837	3988	2238.733	0
	17% ~ 81%	701	1438	2139		
	> 81%	1148	2725	3873		
비정상 동작 처리 하이퍼링크 코드 비율	< 31%	4115	1979	6094	3602.756	0
	31% ~ 67%	767	186	953		
	> 67%	118	2835	2953		
전체		5000	5000	10000		

2.3.2 연속형 변수

상관분석은 변수 간 어떤 선형적 관계를 갖는지 분석하는 기법으로, 상관계수를 이용하여 측정한다. 상관계수 r의 범위는 $-1 \leq r \leq 1$로, $r > 0$인 경우에 양의 상관관계를 가지며, $r < 0$인 경우에는 음의 상관관계를 갖는다. <그림 2.3.3>은 5가지의 연속형 변수 간의 상관관계를 나타낸 것으로, 대부분의 변수들의 상관성이 낮음을 알 수 있다. 비정상 링크 하이퍼링크 코드 비율과 하이퍼링크 코드 비율 두 변수의 상관계수는 -0.24로, 약한 음의 상관관계를 갖는다.



<그림 2.3.3> 연속형 변수 상관관계

3. 머신러닝 모델링

3.1 분류 모형 평가 방법

모델의 실전에서의 성능을 판단하기 위해 6000개의 학습용 데이터 세트와 4000개의 테스트용 데이터 세트로 분할하였다. 총 5개의 기법(로지스틱 회귀, 서포트벡터 머신, 의사결정나무, 랜덤포레스트, 인공신경망)을 사용하여 데이터를 분석하였다. 각 방법에서 나온 최적 모형들을 <표 3.1>의 정확도, 민감도, 특이도, 정밀도, AUC, 이익, 손실, 순이익을 평가하여 가장 최적의 모델을 선택한다. <표 3.1>은 혼동 행렬에서 얻을 수 있는 값으로, 모델을 평가하는 기준이 될 것이다.

순이익은 이익에서 손실을 뺀 값이고, 이익은 피싱을 당하지 않고 예방했을 경우에 피해를 받지 않을 것을 식으로 나타낸 것이다. 피싱을 예방하기 위해서는 실제로 피싱인 웹사이트의 주소가 피싱으로 정확하게 분류되는 것이 가장 중요하다. 따라서 정밀도와 민감도에 가중치를 40%씩 주었다. 그리고 전체적으로 올바르게 예측하는 것도 중요하므로 정확도에 가중치 20%를 주었다.

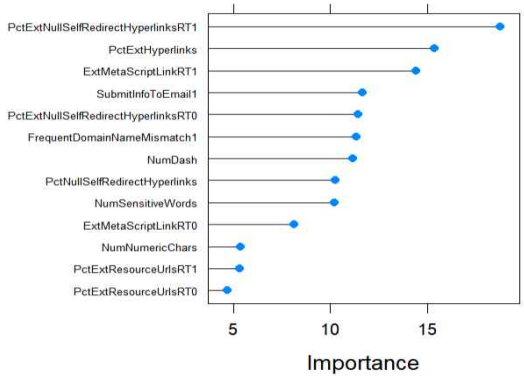
손실은 피싱을 당했을 경우에 발생하는 피해를 식으로 나타낸 것이다. 손실은 잘못 분류하는 경우에 발생하는 것으로 FNR(False Negative Rate)은 실제로 피싱인데 합법인 사이트로 분류한 경우를, FPR(False Positive Rate)은 실제로 합법 사이트인데 피싱인 것으로 분류한 경우다. 우선 전체적으로 잘못 분류하게 되어 발생하는 error의 비율은 (1-정확도)로 가장 중요하므로 가중치 50%를 주었다. 그리고 실제로 피싱인 것을 합법으로 잘못 분류했을 때 오는 손실이 더 클 것이라고 가정해서 FNR은 40%, FPR은 10%로 FNR에 더 높은 가중치를 주었다. FNR은 실제로 피싱인데 합법인 사이트로 분류하는 비율로 (1-민감도)와 같고, FPR은 실제로 합법인데 피싱인 사이트로 분류하는 비율로 (1-특이도)와 같다.

<표 3.1> 분류 모형 평가 방법

정확도	전체 예측 결과 중 올바르게 예측한 비율
민감도	실제 피싱인 데이터 중 피싱으로 예측한 비율
특이도	실제 합법인 데이터 중 합법으로 예측한 비율
정밀도	피싱으로 예측한 결과 중 실제 피싱인 비율
AUC	ROC 곡선 아래 면적
이익	0.4×정밀도+0.4×민감도+0.2×정확도
손실	0.5×(1−정확도)+0.4×(1−민감도)+0.1×(1−특이도)
순이익	이익−손실

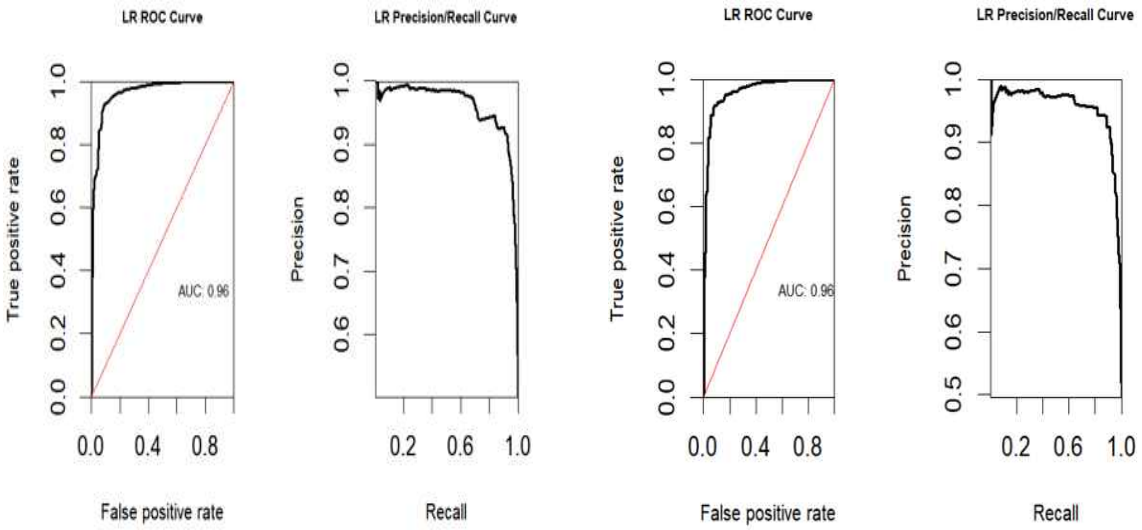
3.2 로지스틱 회귀

로지스틱 회귀(logistic regression)는 종속변수가 범주형인 경우의 회귀 분석으로 각 범주에 속할 확률을 예측하고 분류할 수 있는 통계 기법이다. (Cox, D. R., 1958) 로지스틱 회귀 알고리즘을 이용하여 변수의 중요도를 비교하고, 중요도가 큰 변수들을 선정한다. <그림 3.2.1>은 큰 중요도 값을 갖는 변수 순으로 정렬되어있는 중요도 그래프이다. 비정상 동작 처리 하이퍼링크 코드 비율과 하이퍼링크 코드 비율의 중요도가 가장 높다.



<그림 3.2.1> 로지스틱 회귀의 변수 중요도 그래프

변수 중요도가 10 이하인 문자형 숫자 수와 외부 리소스 코드 비율을 제외하고 나머지 8개 변수만을 고려한 축소 모형을 적합한 후, 전체 모형과 5가지 평가 지표를 비교하여 더 좋은 모형을 선정한다. <그림 3.2.2>는 왼쪽부터 전체 모형의 ROC 곡선과 축소 모형의 ROC 곡선을 비교한 그림이며, 두 모형의 AUC 값은 0.96이다. <표 3.2>는 전체 모형과 축소 모형의 정확도, 민감도, 특이도, 정밀도 그리고 AUC 평가 지표를 정리한 표이다. 정확도, 민감도, 특이도, 정밀도 모두 0.9 이상으로 모두 좋은 모형이다. 정확도에 큰 차이가 없으면서, 효율성을 높이기 위해 축소 모형을 로지스틱 회귀의 최종 모형으로 선정한다.



<그림 3.2.2> 전체 모형과 축소 모형의 ROC 그림

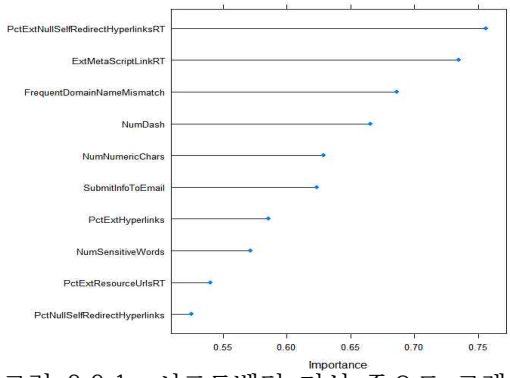
<표 3.2> 로지스틱 회귀 분류 모형 평가 방법

	전체 모형	축소 모형
정확도	0.914	0.910
민감도	0.925	0.917
특이도	0.903	0.902
정밀도	0.904	0.902
AUC	0.96	0.96

3.3 서포트벡터 머신

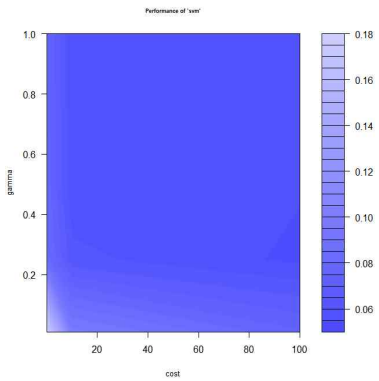
SVM(서포트벡터 머신)은 근접이웃과 회귀방법의 개념을 결합한 모형으로, 초평면이라는 공간을 만들고, 평면의 경계를 만들어서 어떠한 공간에서 동질성을 갖는 그룹들로 분류한다. 주어진 데이터 포인트가 어느 그룹에 속할지 판단하는 이진 선형 분류 모형이다. 정확도 측면에서 다른 분류 기법보다 우수한 결과를 주는 편이나, 결과 해석 측면에서는 직관적이지 않다. (김재희, 2022)

<그림 3.3.1>은 큰 중요도 값을 갖는 변수 순으로 정렬돼있는 중요도 그래프이다. 중요도 순으로 상위 8개 변수를 선택한 모형을 최적화시켜 전체 모형과 5가지 평가 지표를 비교하여 더 좋은 모형을 선정한다.



<그림 3.3.1> 서포트벡터 머신 중요도 그래프

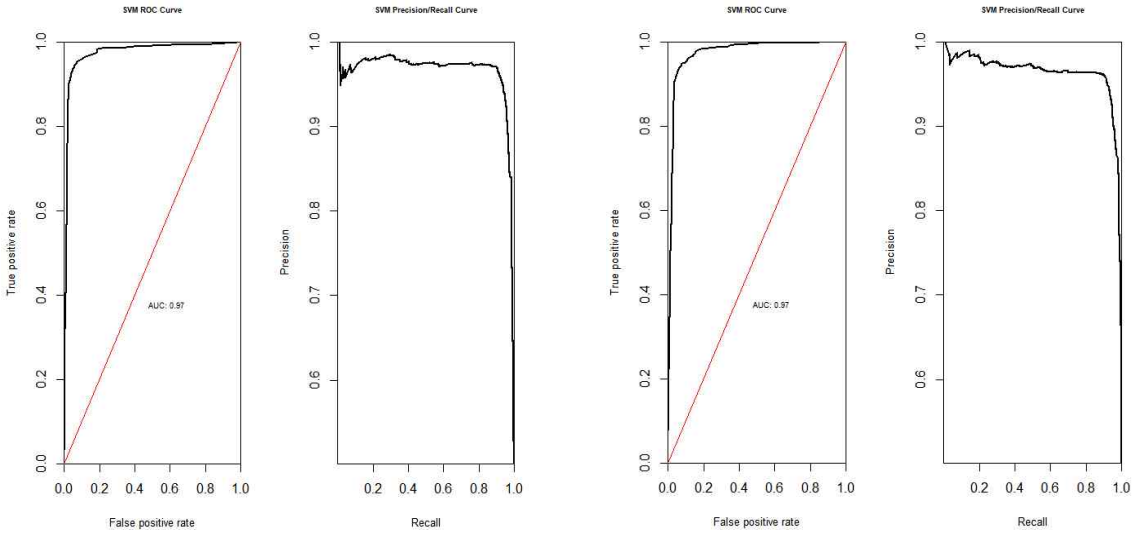
중요도 순으로 상위 8개 변수를 선택한 모형을 다양한 파라미터로 적합하기 위해 gamma는 0.01, 0.25, 0.5, 1로, cost는 0.1, 10, 100로 설정하여 추가적으로 모형을 적합하였다. <그림 3.3.2>를 보면, 튜닝을 한 결과 gamma가 0.25일 때, cost가 100일 때 가장 최적 파라미터 결과가 나왔다. 중요도 순으로 상위 8개 변수를 선택한 모형을 튜닝한 것을 축소 모형으로 정했다.



<그림 3.3.2> 축소 모형 cost&gamma 그래프

전체 모형과 축소 모형을 적합했을 때, <그림 3.3.3>에서 보이는 것처럼 전체 모형의 AUC값

은 0.97, 축소 모형의 AUC값은 0.97로 두 모형 모두 높은 값을 보인다. <표 3.3>은 평가 기준인 정확도, 민감도, 특이도, 정밀도 그리고 AUC 값을 정리한 것으로 전체 모형의 전반적인 값이 더 좋지만, 효율적인 측면에서 축소 모형을 SVM의 최종 모형을 선택하였다.



<그림 3.3.3> 전체 모형과 축소 모형의 ROC 그림

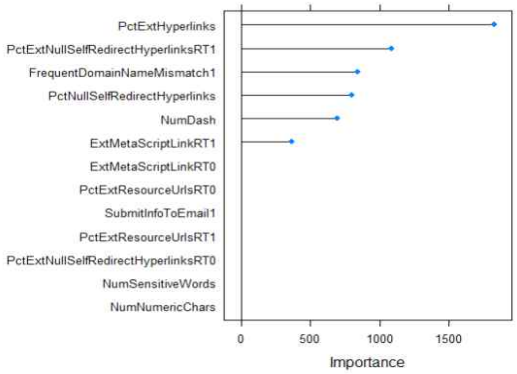
<표 3.3> 서포트벡터 머신 분류 모형 평가 방법

	전체 모형	축소 모형
정확도	0.942	0.936
민감도	0.947	0.942
특이도	0.937	0.930
정밀도	0.938	0.931
AUC	0.97	0.97

3.4 의사결정나무

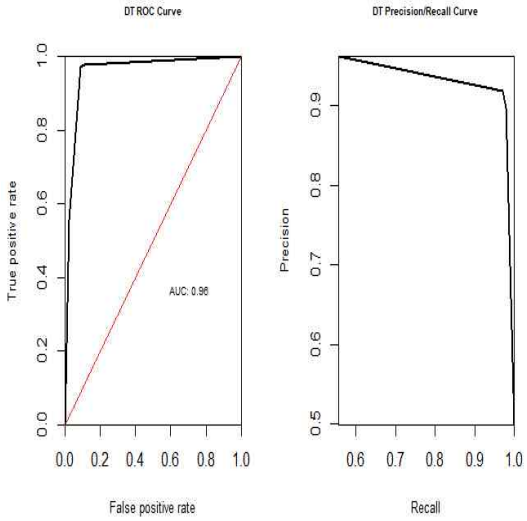
의사결정나무(decision tree)는 나무 구조를 따르는 분류 모형으로 나무를 거꾸로 세운 모양이다. 뿌리 마디(root node)에서 출발하여 결정마디(decision node)를 거쳐 각각의 가지(branch)로 나뉘고, 끝마디(terminal node)에서 의사결정이 마무리된다. 의사결정나무 모형의 시작은 맨 위의 뿌리 마디이며 끝마디까지 분기가 되는 구조이다. 분기 기준(splitting criterion)에 의해 분기가 진행되는데, 분기 기준은 어떠한 독립변수가 종속변수를 잘 구분할 수 있는지를 결정하는 것으로 의사결정의 중요한 구조적 핵심이다. 의사결정나무가 결정되면 모든 끝마디의 데이터 개수를 합한 것과 뿌리 마디의 데이터 개수는 일치하게 된다. 즉 끝마디의 개수가 분리된 집합의 개수가 된다. (김재희, 2022)

의사결정나무 기법을 통해 10개 변수의 중요도를 계산하고, 이를 그래프로 나타낸 것이 <그림 3.4.1>이다. 중요도가 가장 높은 변수 6개는 '하이퍼링크 코드 비율', '비정상 동작처리 하이퍼링크 코드 비율', '도메인 이름 불일치', '비정상링크 하이퍼링크 코드 비율', '대시 기호 수', '외부 파일 코드 비율'인 것을 확인할 수 있다. 상위 6개 변수로 모형을 적합하여 새로운 축소 모형을 만들었다.



<그림 3.4.1> 의사결정나무 중요도 그래프

<그림 3.4.2>는 전체 모형의 ROC 곡선과 축소 모형의 ROC 곡선의 그림이고, <표 3.4.3>는 전체 모형과 축소 모형의 정확도, 민감도, 특이도, 정밀도, AUC를 나타낸 것이다. 전체 모형과 축소 모형의 결과는 차이가 없었고, 변수의 수가 적어서 더 효율적인 축소 모형을 최종 모형으로 선택한다.

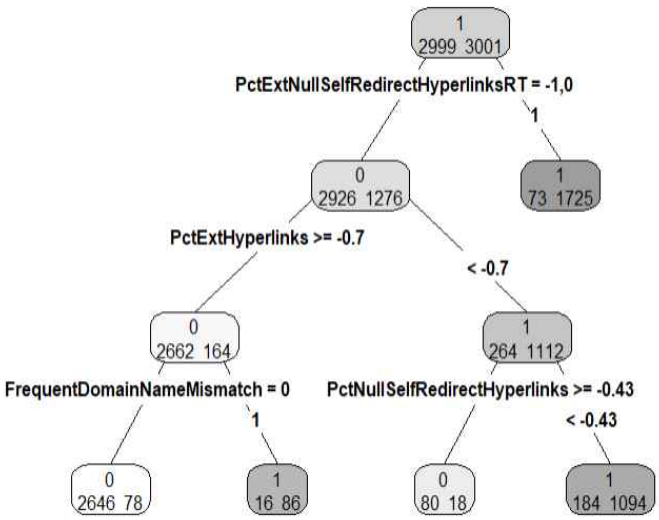


<그림 3.4.2> 전체 모형과 축소 모형의 ROC 그림

<표 3.4.3> 의사결정나무 분류 모형 평가 방법

	전체 모형 & 축소 모형
정확도	0.942
민감도	0.971
특이도	0.913
정밀도	0.91
AUC	0.96

<그림 3.4.3>은 최종 모형의 의사결정나무를 시각화한 것이다. 뿌리노드에서 합법 웹사이트이면 왼쪽 노드로, 피싱 웹사이트이면 오른쪽 노드로 분기되는데, 가장 먼저 '비정상 동작처리 하이퍼링크 코드의 비율'이 67% 초과이면 오른쪽 노드로, 67% 이하이면 오른쪽 노드로 분기됐다. 다음 결정 노드에서 '하이퍼링크 코드의 비율'이 표준화된 값이 -0.7 이상이면 왼쪽 노드로, -0.7 미만이면 오른쪽 노드로 분기되었다. 위의 결정 노드에서 합법 웹사이트로 분기된 후 '도메인 이름 불일치'가 일치이면 왼쪽 노드, 불일치이면 오른쪽 노드로 분기되었고, 반대로 위의 결정 노드에서 피싱 웹사이트로 분기된 후 '비정상링크 하이퍼링크 코드 비율'의 표준화된 값이 -0.43 이상이면 왼쪽 노드, -0.43 미만이면 오른쪽 노드로 분기됐다. 총 5개의 끝마디로 의사결정이 마무리되었다.

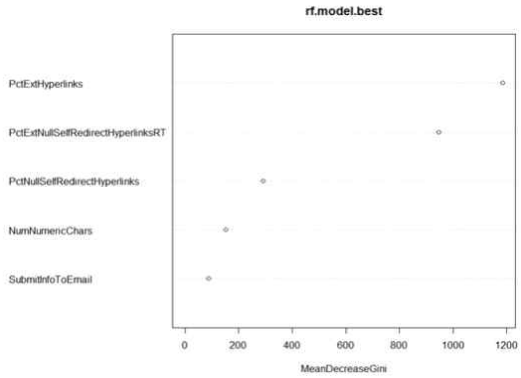


<그림 3.4.3> 의사결정나무 시각화

3.5 랜덤포레스트

랜덤포레스트는 데이터를 반복 복원 추출하고, 거기에 더해 변수도 랜덤하게 추출하여 만들어진 다양한 나무 모형의 앙상블이다. 데이터를 반복 복원 추출하고 각 데이터마다 학습을 수행하여 투표방식으로 결과를 집계하는 방법인 배깅(Bagging)을 진행하면 비슷한 나무 모형이 만들어질 가능성이 커지며 전체 모델의 분산이 증가하는 결과를 초래할 수도 있다. 랜덤포레스트는 배깅을 진행하고 거기에 더해 변수도 랜덤하게 추출하기 때문에 전체 모델의 분산을 감소시키므로 모형의 성능이 개선된다는 장점을 가진다. (김재희, 2022)

변수 10개를 모두 넣은 전체 모형과 파라미터 최적화를 진행한 축소 모형을 비교한다. 축소 모형은 Feature Selection 알고리즘을 통해 선택된 변수를 적합한 모형의 하이퍼 파라미터를 조정하여 파라미터 최적화를 진행하였다. 축소 모형을 비교한다. <그림 3.5.1>은 랜덤포레스트의 Feature Selection으로 선택된 변수의 중요도 그래프이다.



<그림 3.5.1> 랜덤포레스트 변수 중요도 그래프

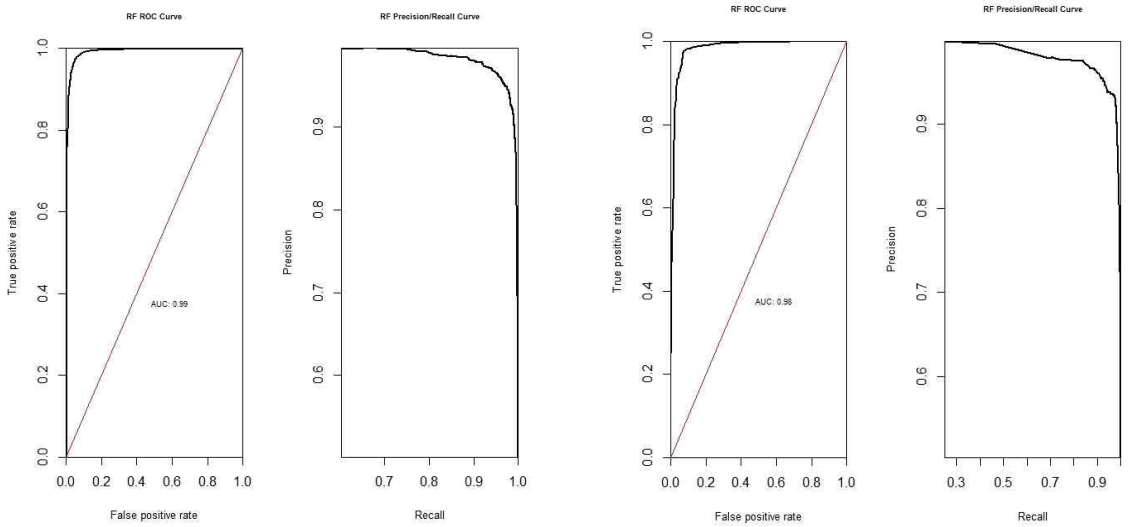
<표 3.5.1>은 전체 모형의 혼동행렬로 잘못 예측한 개수가 163개이고, <표 3.5.2>는 축소 모형의 혼동행렬로 잘못 예측한 개수가 192개이다. 이를 통해 축소 모형이 전체 모형보다 잘못 예측한 개수가 더 많음을 알 수 있다. <그림 3.5.2>은 전체 모형과 축소 모형의 ROC 그림으로 AUC값은 각 0.99와 0.98로 높은 적합률을 보인다. Feature Selection 방법에서 선택된 변수는 '하이퍼링크 코드 비율', '메일 보내기 기능 존재', '비정상링크 하이퍼링크 코드 비율', '비정상 동작 처리 하이퍼링크 코드 비율', '문자형 숫자 수'이며, 노드 사이즈가 2, 후보군이 3, 트리의 개수가 200일 때 최적의 파라미터의 모형이 나왔다.

<표 3.5.1> 전체 모형 혼동행렬

		Actual class	
		0	1
Predicted class	0	1887	66
	1	97	1950

<표 3.5.2> 축소 모형 혼동행렬

		Actual class	
		0	1
Predicted class	0	1843	51
	1	141	1965



<그림 3.5.2> 전체 모형과 축소 모형의 ROC 그림

두 모형 모두 정확도와 민감도가 높지만, 효율성을 위해 Feature Selection 방법으로 선택한 5개의 변수를 파라미터를 튜닝한 축소 모형을 랜덤포레스트의 최종 모형으로 선택하였다.

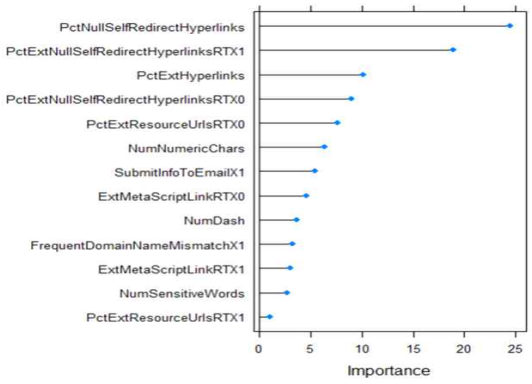
<표 3.5.3> 랜덤포레스트 분류 모형 평가 방법

	전체 모형	축소 모형
정확도	0.9592	0.952
민감도	0.9673	0.9747
특이도	0.9511	0.9289
정밀도	0.9526	0.933
AUC	0.99	0.98

3.6 인공신경망

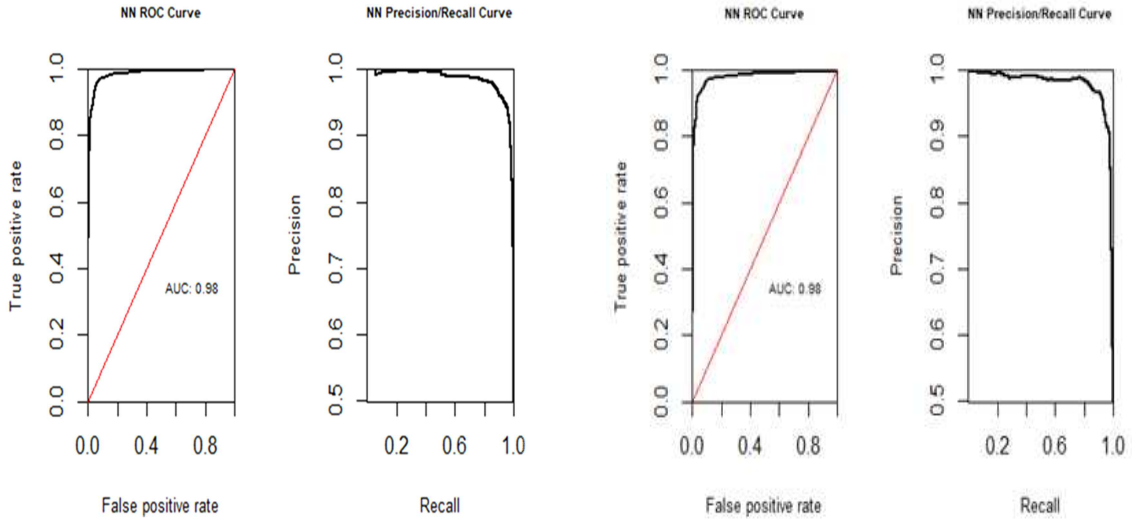
인공신경망이란 데이터의 입력과 출력의 관계를 생물학적으로 뇌가 어떤 자극 때문에 반응하는 것과 같은 개념에서 파생된 모형이다. 뇌에서는 뉴런을 통해 자극을 전달한다면, ANN 모형에서는 노드를 통해 데이터 정보를 전달한다. ANN 모형은 활성화 함수를 통해 입력을 출력으로 변환한다. (김재희, 「다변량 및 빅데이터 분석 실무」, 2022, p.239)

<그림 3.6.1>은 인공신경망 전체 모형으로 도출한 변수 중요도 그래프이다. '비정상링크 하이퍼링크 코드 비율'과 '비정상 동작 처리 하이퍼링크 코드 비율'의 중요도가 가장 높았으며, '민감한 단어 수'와 '도메인 이름 불일치'의 중요도가 다른 변수에 비해 낮았다. 변수 중요도가 5 이하인 '민감한 단어 수', '도메인 이름 불일치'를 제외한 나머지 상위 8개 변수로 축소 모형1을 적합하고, 상위 7개 변수로 축소 모형2를 적합했다.

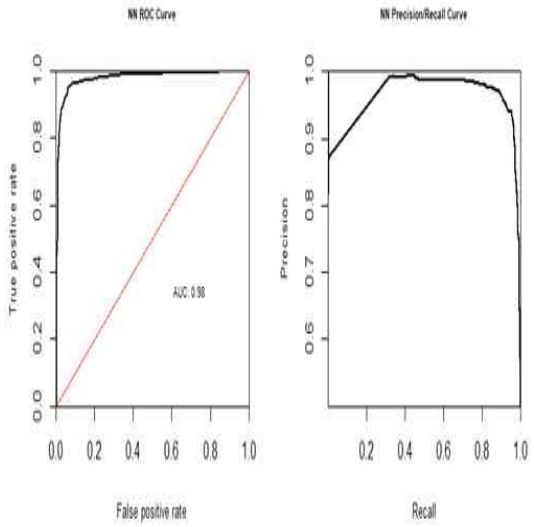


<그림 3.6.1> 인공신경망 중요도 그래프

먼저 10개 변수를 넣은 전체 모형을 적합했을 때, <그림 3.6.2>과 <그림 3.6.3>에서 보이는 것처럼, 전체 모형의 AUC값은 0.98, 축소 모형1의 AUC값은 0.98, 축소 모형2의 AUC값은 0.98로 전부 높은 적합률을 보인다.



<그림 3.6.2> 전체 모형과 축소 모형1의 ROC 그림



<그림 3.6.3> 축소 모형2의 ROC 그림

각 모형의 성능을 평가한 <표 3.6>을 보면, 전반적으로 모두 성능이 높다. 전체 모형의 성능이 가장 우수하지만, 변수가 7개인 축소 모형2의 성능이 전체 모형과 매우 유사하다. 따라서 효율성을 고려하여 변수가 7개인 축소 모형2를 인공신경망의 최종 모형으로 선택한다. 최종 모형으로 선택된 변수는 '비정상링크 하이퍼링크 코드 비율', '비정상 동작 처리 하이퍼링크 코드 비율', '하이퍼링크 코드 비율', '외부 리소스 코드 비율', '숫자형 문자 개수', '메일 보내기 기능 존재', '외부 파일 코드 비율'순으로 중요하다. 이를 통해 하이퍼링크와 관련된 변수들이 이 중요도가 높다고 판단됨을 알 수 있다.

<표 3.6> 인공신경망 분류 모형 평가 방법

	전체 모형	축소 모형1 (8개)	축소 모형2 (7개)
정확도	0.949	0.941	0.941
민감도	0.957	0.931	0.943
특이도	0.942	0.949	0.939
정밀도	0.942	0.950	0.938
AUC	0.98	0.98	0.98

4. 결과

<표 4.1>의 5가지 모델링의 분류 평가 방법을 통해 모두 총합하여 랜덤포레스트 모형을 최종 모형으로 선택하였다. 최종 모형은 Feature Selection 방법으로 선택된 변수 '하이퍼링크 코드 비율', '메일 보내기 기능 존재', '비정상 링크 하이퍼링크 코드 비율', '비정상 동작 처리 하이퍼링크 코드 비율', '문자형 숫자 수'로 만든 모형의 파라미터를 튜닝한 랜덤포레스트 모형이다.

랜덤포레스트가 모든 분류에서 98%로 가장 좋은 적합률을 가진다. 전체 예측 결과 중 올바르게 예측한 비율이 가장 높으며, 0에 가장 가까운 손실 값을 가진다. 또한 랜덤포레스트는 각 모델의 데이터와 특성을 무작위로 구성함으로써 다양한 나무를 생성하고 결합하여 평균을 내는 장점이 있다. 이는 의사결정나무의 과적합을 예방할 수 있으며 더 좋은 정확도를 갖게 한다. 본 분석에서도 랜덤포레스트가 의사결정나무에 비해 약 2% 높은 결괏값을 가진다. 과적합을 예방하기 위해서는 데이터의 양을 늘리는 방법이 있다. 하지만 랜덤포레스트의 특성상 많은 데이터를 모델링 하는데 오랜 시간이 걸리기 때문에 어려움이 있다. 변수를 10개로 감축했음에도 불구하고 1만 개의 데이터를 다루는 데 소요된 시간이 매우 컸다. 가장 위험한 경우인 피싱 웹사이트를 합법 웹사이트로 예측할 확률도 2.5%로 다른 모델에 비해 확연히 낮다.

<표 4.1> 5가지 모형의 분류 모형 평가 방법

	로지스틱 회귀	서포트벡터 머신	의사결정나무	랜덤포레스트	인공신경망
정확도	0.910	0.936	0.942	0.952	0.941
민감도	0.917	0.942	0.971	0.975	0.943
특이도	0.902	0.930	0.913	0.929	0.938
정밀도	0.902	0.931	0.918	0.933	0.939
AUC	0.96	0.97	0.96	0.98	0.98
이익	0.91	0.936	0.944	0.953	0.941
손실	0.088	0.062	0.05	0.041	0.059
순이익	0.822	0.874	0.894	0.912	0.882

5. 결론

본 분석에서 사용한 데이터는 피싱 웹사이트와 합법 웹사이트에서 추출한 48개의 기능을 포함하고 있다. 분석의 효율성을 위해 논문을 참고하여 종속변수를 포함한 11개의 변수를 선택하였고, 연속형 변수와 범주형 변수로 나누어 탐색적 자료 분석을 진행하였다. 범주형 변수와 종속변수의 연관성을 알아보기 위해 카이제곱 검정을 실시하였다. 모든 범주형 변수에 대해 p

-값이 0.05보다 작기 때문에 유의수준 0.05 하에서 각 범주형 독립변수와 종속변수 사이에 연관성이 있다고 판단했다. 따라서 모든 범주형 변수를 분석에 사용하였다. 그리고 연속형 변수끼리의 상관성을 보았을 때, 대부분의 변수들의 상관성이 낮았다. 그중에서도 가장 높은 상관관계를 보이는 변수들은 비정상 링크 하이퍼링크 코드 비율과 하이퍼링크 코드 비율로 -0.24의 약한 음의 상관관계를 보였다.

각 변수들에 대해 살펴본 후, 머신러닝의 지도학습 중에서 로지스틱 회귀, 서포트벡터 머신, 의사결정나무, 랜덤포레스트, 인공신경망을 통해 웹사이트의 피싱 여부를 판단할 수 있는 분류 모형을 만들었다. 우선 각자의 기법으로 10개의 변수를 적용한 전체 모형과 각 중요도 그래프에 의해 선택된 변수들로 이루어진 축소 모형을 비교하여 더 나은 모형을 선택하였다. 총 5개의 모형을 후보로 선택하였고, 이 모형들을 비교하기 위해서 정확도, 민감도, 특이도, 정밀도, AUC, 이익, 손실, 순이익을 모형 평가 기준으로 삼았다.

랜덤포레스트 기법에서 Feature Selection 방법으로 높은 중요도를 갖는 5개의 변수를 찾았고, 그 결과 변수 '하이퍼링크 코드 비율', '메일 보내기 기능 존재', '비정상링크 하이퍼링크 코드 비율', '비정상 동작 처리 하이퍼링크 코드 비율', '문자형 숫자 수'가 선택되었다. 선택된 변수로 이루어진 모형에 노드 사이즈가 2, 후보군이 3, 트리의 개수가 300으로 튜닝하여 랜덤포레스트 기법의 최적의 모형을 찾았다. 그리고 이 모형을 다른 기법의 모형들과 비교했을 때 정확도, AUC, 이익, 손실, 순이익의 측면에서 가장 좋은 값을 가진다. 따라서 랜덤포레스트의 축소 모형을 성능이 가장 좋은 모형이라고 판단하여 최종 모형으로 선택하였다.

범주형 독립변수와 종속변수의 카이제곱 분석, 의사결정나무, 변수 중요도 그래프를 통해서 어떤 변수가 피싱 여부에 영향을 주는지 알 수 있었다. 카이제곱 분석을 통해 도메인 이름이 불일치하는 경우, 비정상 동작 처리 하이퍼링크 코드 비율이 67%를 초과하는 경우의 피싱 비율이 0.9 이상이기 때문에, 두 가지 경우가 피싱 웹사이트의 특징을 잘 나타낸다고 할 수 있다. 또한 의사결정나무에서 ‘비정상 동작처리 하이퍼링크 코드 비율’, ‘하이퍼링크 코드 비율’, ‘도메인 이름 불일치’, ‘비정상링크 하이퍼링크 코드 비율’이 합법과 피싱을 분류할 때 분기기준으로 사용되었다. 각 기법 별 변수 중요도 그래프에서는 ‘하이퍼링크 코드 비율’과 ‘비정상 동작 처리 하이퍼링크 코드 비율’이 5가지 모형의 중요도 그래프에서 선택되었고, '비정상링크 하이퍼링크 코드 비율', ‘외부 파일 코드 비율’, ‘메일 보내기 기능’이 4가지 모형의 중요도 그래프에서 선택되었다.

따라서 본 연구는 랜덤포레스트의 변수 중요도 분석을 통해서 선택된 5개의 변수로 만든 모형을 기반으로 웹사이트의 피싱 여부를 판단할 수 있는 시스템을 도입하는 것을 제안하고자 한다. 그래서 이 시스템이 pc 확장 프로그램이나 모바일 어플리케이션을 통해 널리 사용되면 온라인 이용자들에게 안전한 사이버 환경을 제공해 줄 수 있을 것이라고 기대한다.

참고 문헌

Internet Crime Complaint Center (2021), 「Internet Crime Report 2021」, https://www.ic3.gov/Media/PDF/AnnualReport/2021_IC3Report.pdf

김한경. "과기정통부, '올해 사이버위협 분석과 2022년 사이버위협 전망' 발표", 뉴스투데이, 2021년 12월 27일. https://www.news2day.co.kr/article/20211227500098

Chiew, K. L., Tan, C. L., Wong, K., Yong, K. S., & Tiong, W. K. (2019), 「A new hybrid ensemble feature selection framework for machine learning-based phishing detection system」, Information Sciences, 484, 153-166

Mohammad R. M. A., McCluskey L. & Thabtah F. (2015), UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/phishing+websites, University of Huddersfield&Canadian University of Dubai

Cox, D. R. (1958), 「The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological)」, 20(2), 215-232.

김재희 (2022), 「다변량 및 빅데이터 분석 실무」, 115, 212, 316, 239