그거 맞아?

피싱 사이트 탐지

트릴리언

201620899 안윤회

201620641 유상정

201620631 김상우

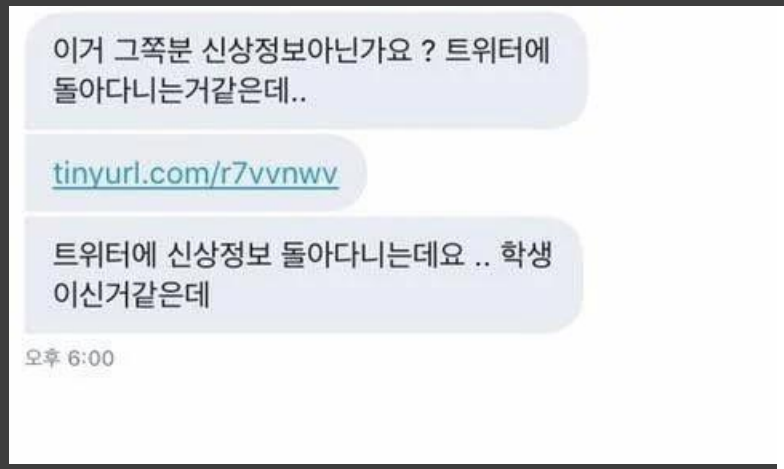
201620630 김두원

201620648 이찬호

Contents

1. Introduction

1-1. Background



피해자는 모두 여중생 3명으로 피싱 사이트를 이용해 피해자를 유인한 뒤 성 착취 영상을 찍은 뒤 협박한 것으로 드러났다.

조사결과 이들은 'n번방'을 모방해 범행에 나선 것으로 알려졌다. '프로젝트 N'이라는 명칭으로 범죄를 모의했다고 알려졌다. 이어 기존 'n번방'의 음란물을 물려받아 재판매해 2500만원의 이익을 챙긴 운영자도 검거했다.

강원경찰은 텔레그램 음란물 유포 방식을 최초 도입한 '와치맨'(38세 전모씨)의 수사도 한 것으로 밝혀졌다.

<최근 4년 간 신고·차단된 피싱사이트 현황>

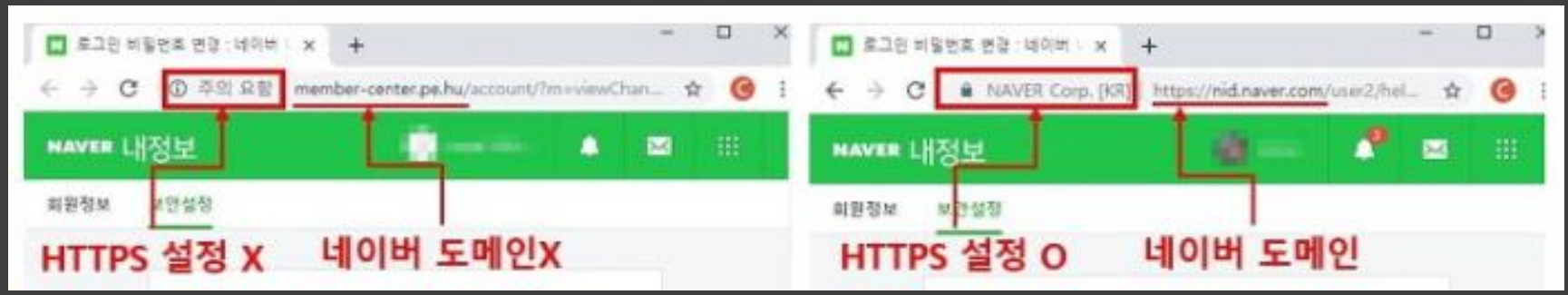
(단위 : 건)

구분	2016	2017	2018	2019.8월	합계
피싱사이트 신고·차단	4,286	10,469	9,522	7,063	31,340

※출처 : 과학기술정보통신부

※피싱의 경우 전자우편 또는 메신저 등을 통해 전달되어 탐지현황 대신 신고현황

1-1. Background



< 정상 금융감독원
e-금융지원센터 사이트 >

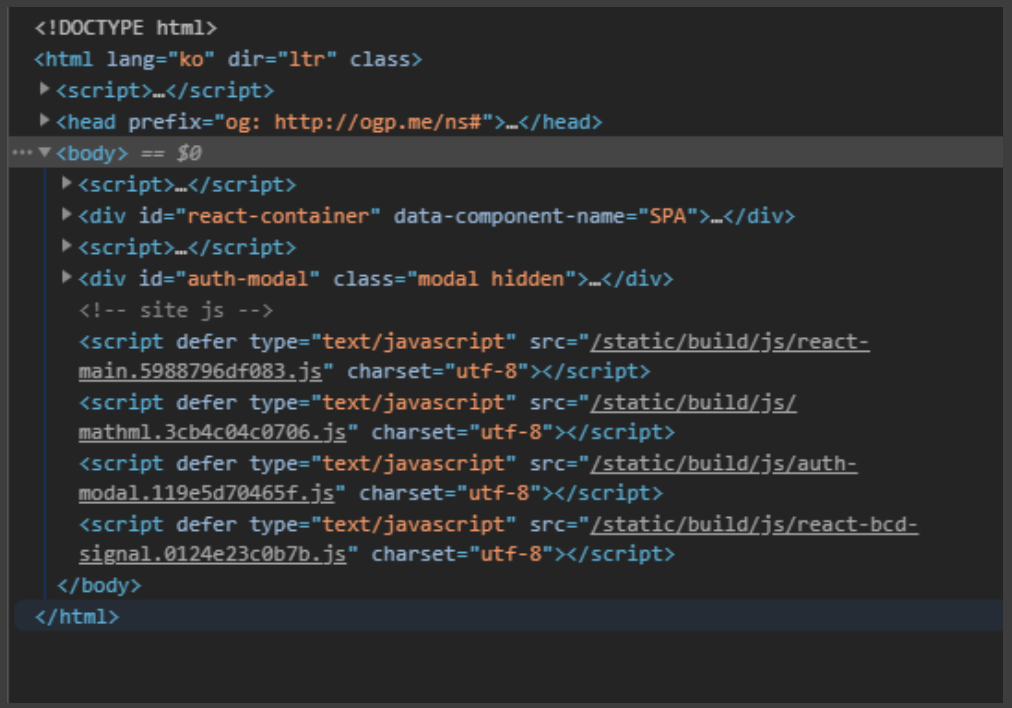
< 가짜 금융감독원
e-금융지원센터 사이트 >

1-1. Background

문제점 : URL의 특징만으로 충분한 성능이 나오지 않음.

개선 방안 :

웹 사이트에 공개되어있는 html과 javascript에 포함된 정보를 토대로 새로운 feature 선정



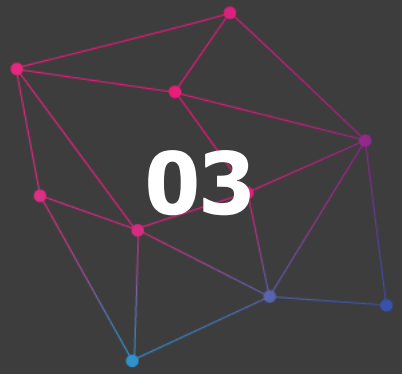
1-2. Project Summary

Subject : 기계학습을 이용한 피싱 사이트의 탐지

프로젝트 목표

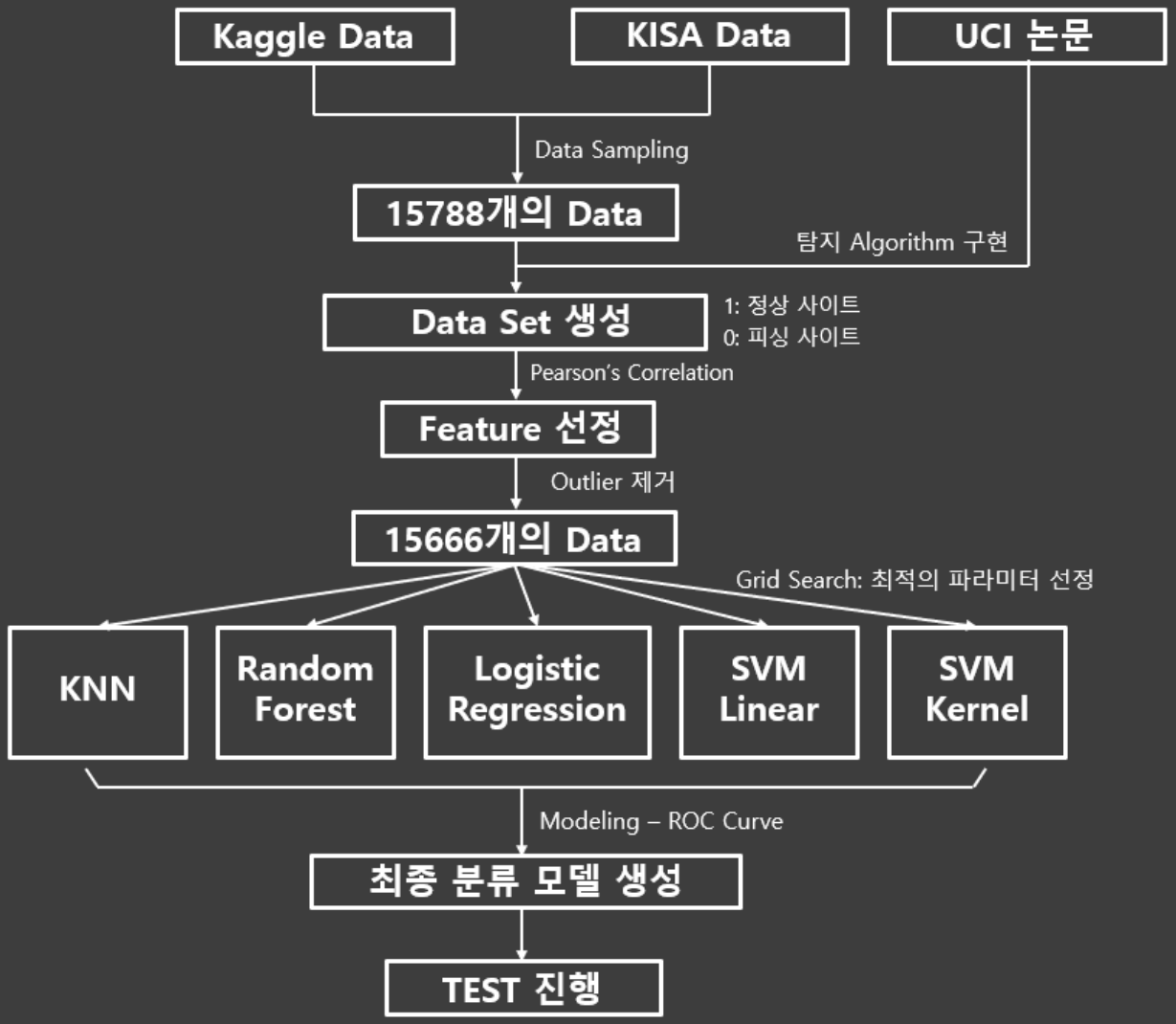
피싱 사이트의 특징 파악
및 Data feature 선정

알고리즘별 평가를 통한
가장 적합한 알고리즘 판단

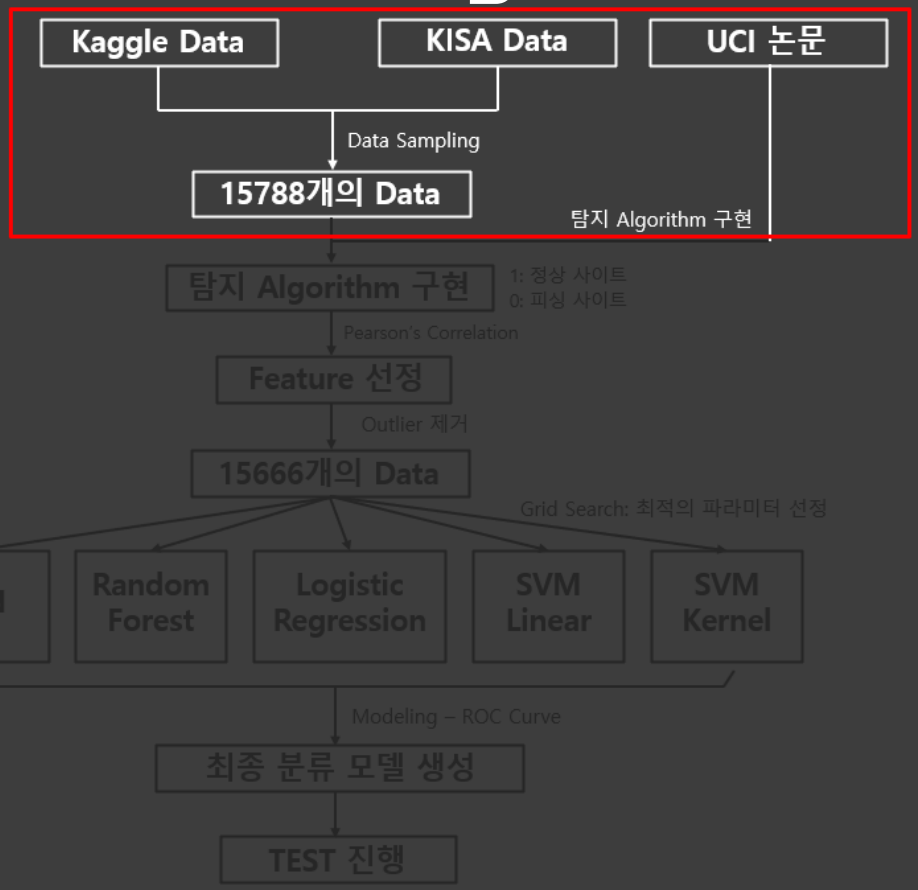


피싱 사이트와
정상사이트 구별

1-2. Project Summary



2. Data Understanding



2-1. Data set

2-2. Phishing site 특징

KAGGLE에서 참고한 URL 기반 Feature

1. Double_slash_redirection : http://을 제외한 //의 존재 여부
2. having_At_Symbol: URL내에 '@'존재 여부
3. Shortening_Service : URL 단축 서비스 사용 여부
4. having_Sub_Domain : 2개 이상의 sub domain 존재 여부
5. Prefix_suffix : URL 내에 '-'존재 여부
6. URL_Length: 길이가 74 이상인 URL 구별
7. Favicon: Favicion 존재 여부 및 외부 도메인에서 Favicon 참조 여부

2-2. Phishing site 특징

KAGGLE에서 참고한 contents기반 Features

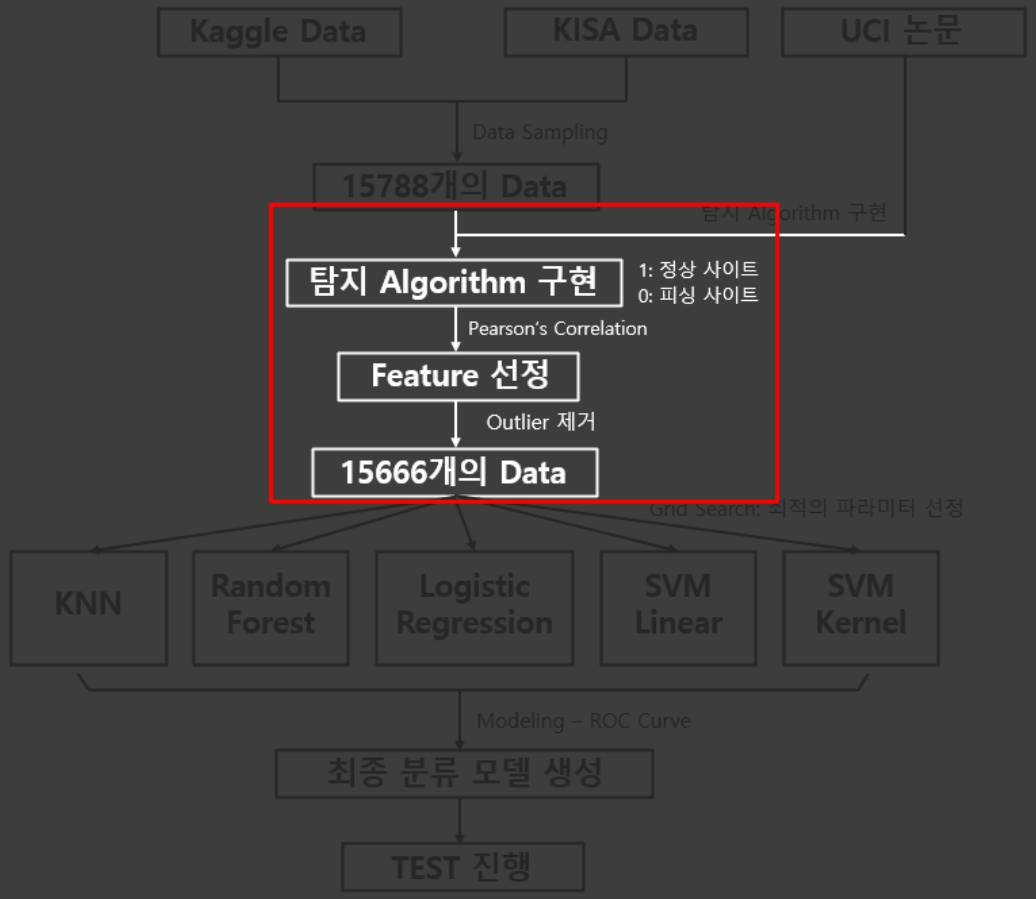
1. Duplicated_Tag : 소스코드 내에 <Head>나 <body> tag의 2번 이상 등장 여부
2. Count_Redirection: 3회 이상의 Redirection 여부
3. getAnchorResult: <a> tag의 67%이상이 외부 도메인으로 연결 여부
4. getLinksInTags: 모든 tag 중<meta>, <script>, <link> tag가 50% 이상인지 여부
5. Prompt_in_popup: 사용자에게 입력을 요구하는 팝업의 존재 여부
6. Disabling_Right_Click : 마우스 우클릭 가능 여부
7. Domain_registration_length : 도메인 가입 기간이 1년 이하인지 여부
8. Alexa_Ranking : 알렉사 랭킹 포함 여부
9. DomainName_in_Source : 소스코드 내에 자기 도메인 이름 포함 여부

2-2. Phishing site 특징

논문을 기반으로 새로 고려해본 feature

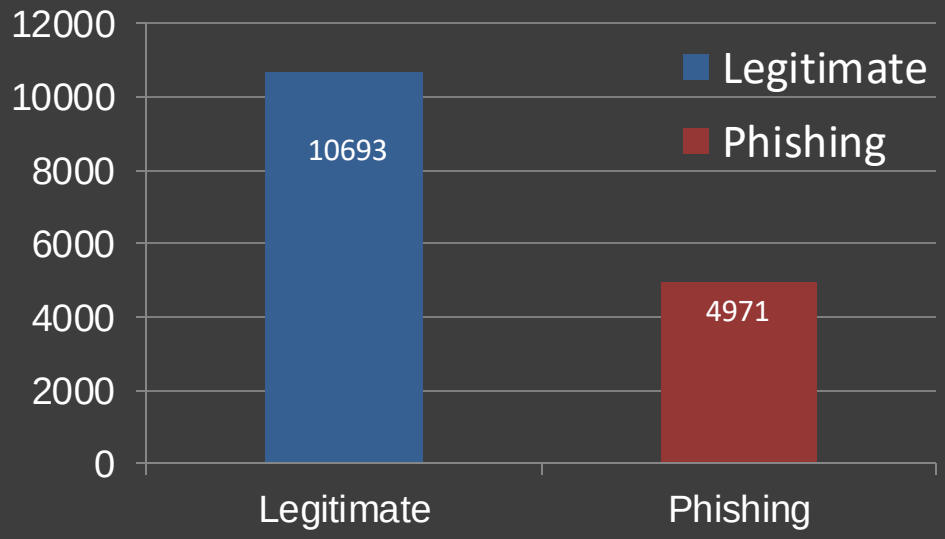
1. AbnormalUrl : Whois DB에 도메인 존재 여부
2. Form_action_Handler: 소스 코드의 action 속성에 공백이나 about:blank 여부
3. Remain_Expiration : 도메인 유효기간이 6달 보다 짧은 지 여부
4. onMouseOver : 상태 표시줄에서의 OnMouseOver기능 사용 여부
5. Length_of_Source : 소스코드의 길이가 50000 이상인지 여부
6. External_Load_Script : script가 100% 외부에서 로드 되었는지 여부

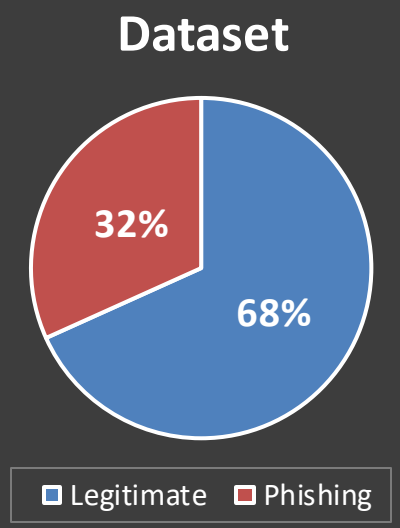
3. Data preparation (Pre-processing)



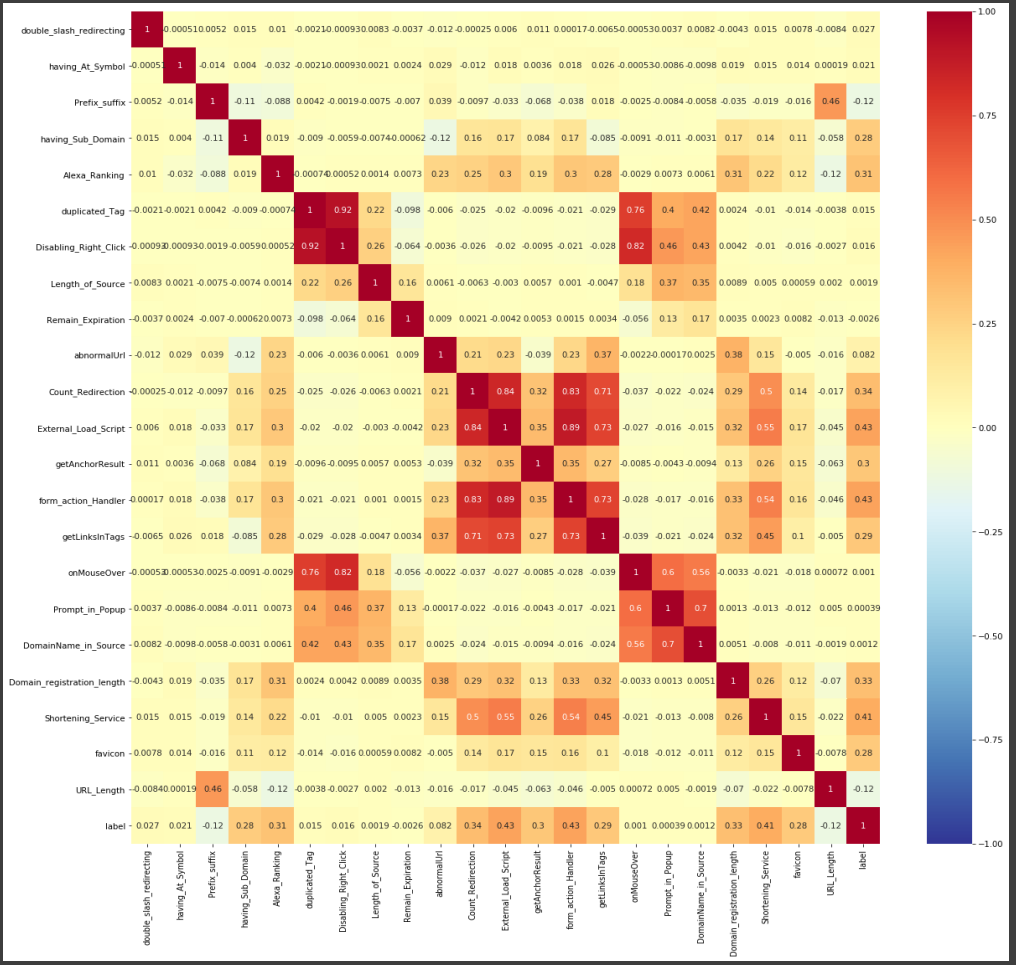
3-1. Data Cleaning

1. 총 수집한 42만개의 데이터 셋 중 겹치는 도메인을 제거한 15만개 선택. (중복 data 제거)
2. 15만개 중, 15788개의 Dataset 무작위로 선별. (Data Sampling)
3. 정상 URL 중, contents 기반 featur가 모두 0일 경우 피싱 사이트로 분류될 수 있기 때문에 제거 -> 15666개의 Data set 사용. (Outlier 제거)





3-2. Feature 선정

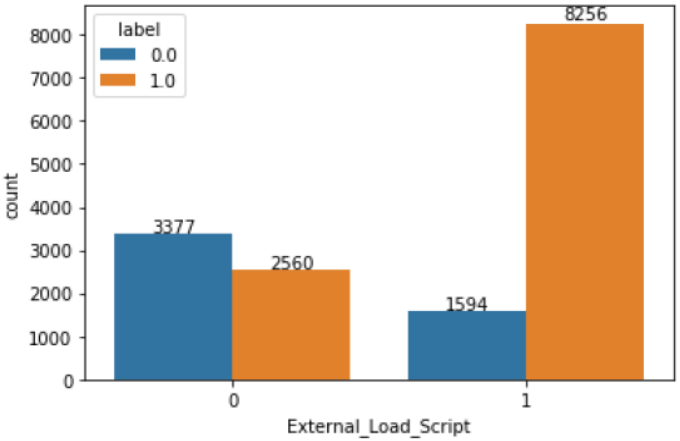


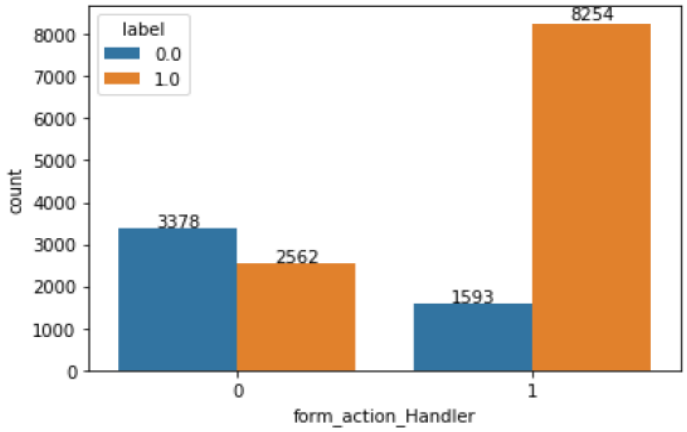
Pearson's Correlation
0.1 이하인 Feature 삭제

```
del df['double_slash_redirecting']
del df['having_At_Symbol']
del df['duplicated_Tag']
del df['Disabling_Right_Click']
del df['Length_of_Source']
del df['Remain_Expiration']
del df['abnormalUrl']
del df['onMouseOver']
del df['Prompt_in_Popup']
del df['DomainName_in_Source']
```

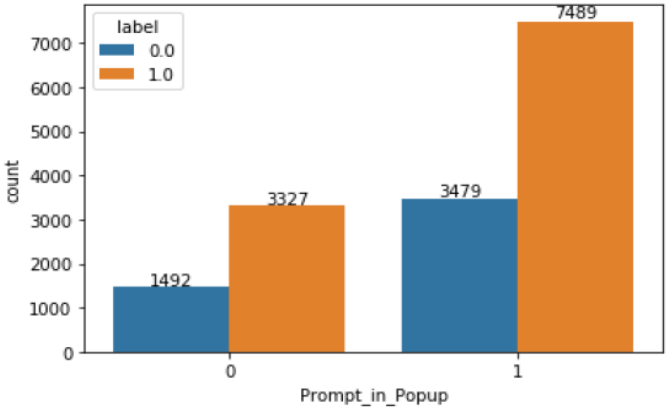
3-2. Feature 선정

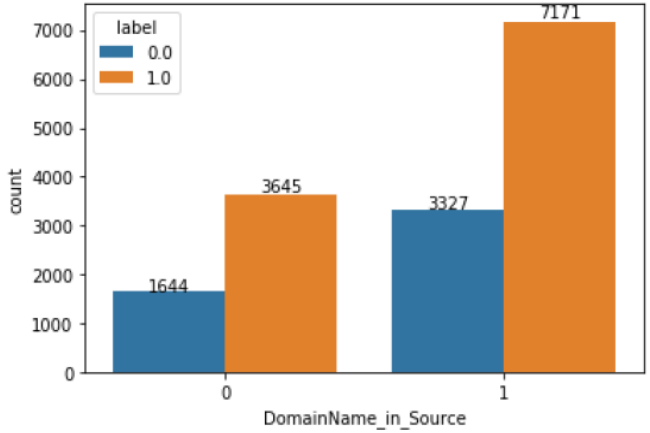
High correlation





Low correlation





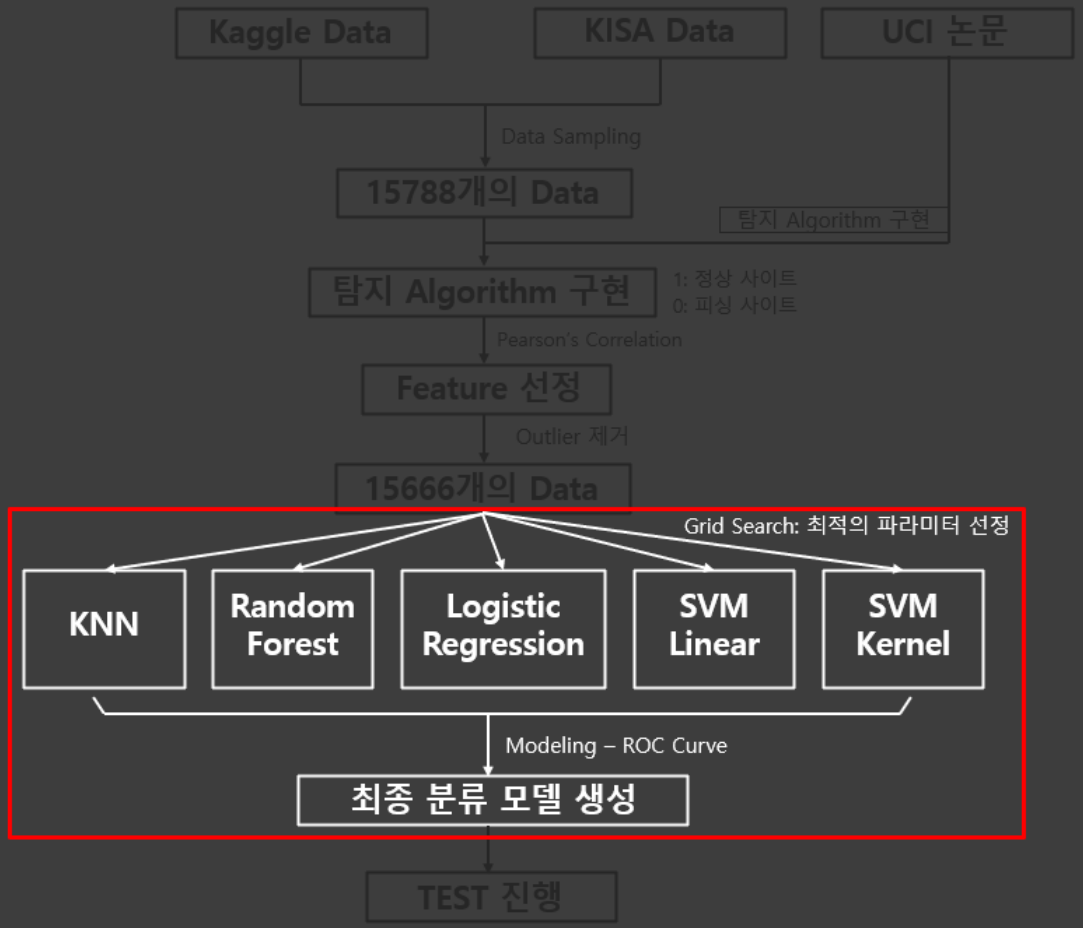
3-3. 새롭게 생성한 dataset

	having_Sub_Domain	Alexa_Ranking	Count_Redirection	External_Load_Script	getAnchorResult	form_action_Handler	getLinksInTags	Domain_registration_length	Shortening_Service	favicon
0	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	0	1	1	1	1	1
3	1	1	0	0	0	0	1	1	0	0
4	1	1	1	1	0	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...
15783	1	0	1	1	0	1	1	1	1	0
15784	1	0	1	1	0	1	1	1	1	0
15785	1	0	0	0	0	0	0	0	1	1
15786	1	1	1	1	1	1	1	1	1	0
15787	1	0	1	1	0	1	1	0	1	1

15666 rows × 10 columns

```
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Prefix_suffix               15666 non-null  int64
 1   having_Sub_Domain           15666 non-null  int64
 2   Alexa_Ranking               15666 non-null  int64
 3   Count_Redirection           15666 non-null  int64
 4   External_Load_Script        15666 non-null  int64
 5   getAnchorResult             15666 non-null  int64
 6   form_action_Handler         15666 non-null  int64
 7   getLinksInTags              15666 non-null  int64
 8   Domain_registration_length  15666 non-null  int64
 9   Shortening_Service          15666 non-null  int64
 10  favicon                     15666 non-null  int64
 11  URL_Length                  15666 non-null  int64
dtypes: int64(12)
```

4. Modeling


Kaggle Data
KISA Data
UCI 논문
Data Sampling
15788개의 Data
탐지 Algorithm 구현
탐지 Algorithm 구현
1: 정상 사이트
0: 피싱 사이트
Pearson's Correlation
Feature 선정
Outlier 제거
15666개의 Data
Grid Search: 최적의 파라미터 선정
KNN
Random Forest
Logistic Regression
SVM Linear
SVM Kernel
Modeling – ROC Curve
최종 분류 모델 생성
TEST 진행

4. Modeling

평가 지표 → ROC 곡선
AUC 값

평가 지표 선택 이유

Accuracy → Data 에 따라서 달라질 수 있다.

4. Grid Search

Hyperparameter tuning 과정

```
from sklearn.svm import SVC # SVM classifier 사용위한 import
from sklearn.model_selection import GridSearchCV

# parameter 정의
param_grid = {'C': [1, 0.1, 0.01, 0.001, 0.0001],
              'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
              'kernel': ['linear']}
#SVM grid Search
SVC_linear_grid = GridSearchCV(SVC(), param_grid, refit = True, verbose = 3, n_jobs = 8).fit(x_train,y_train.values.ravel())
```

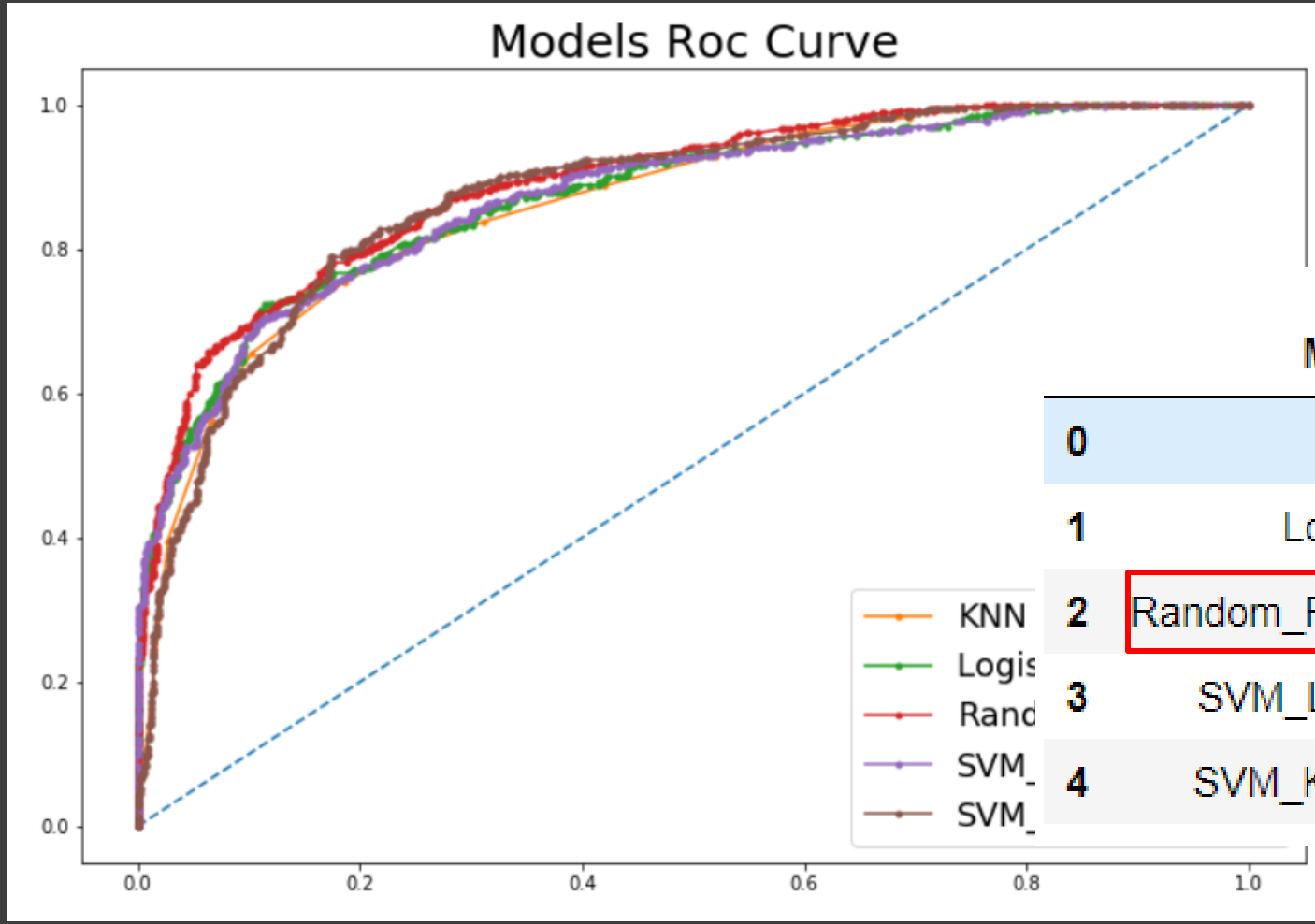
```
#model evaluation
print("test set score : {}".format(SVC_linear_grid.score(x_test,y_test)))

#best parameters and best score
print("best parameters : {}".format(SVC_linear_grid.best_params_))
print("best score : {}".format(SVC_linear_grid.best_score_))
```

```
test set score : 0.799617102744097
best parameters : {'C': 1, 'gamma': 1, 'kernel': 'linear'}
best score : 0.8194201203082427
```

예시 : SVM (linear)

4. Modeling

모든 feature가 존재하는 dataset

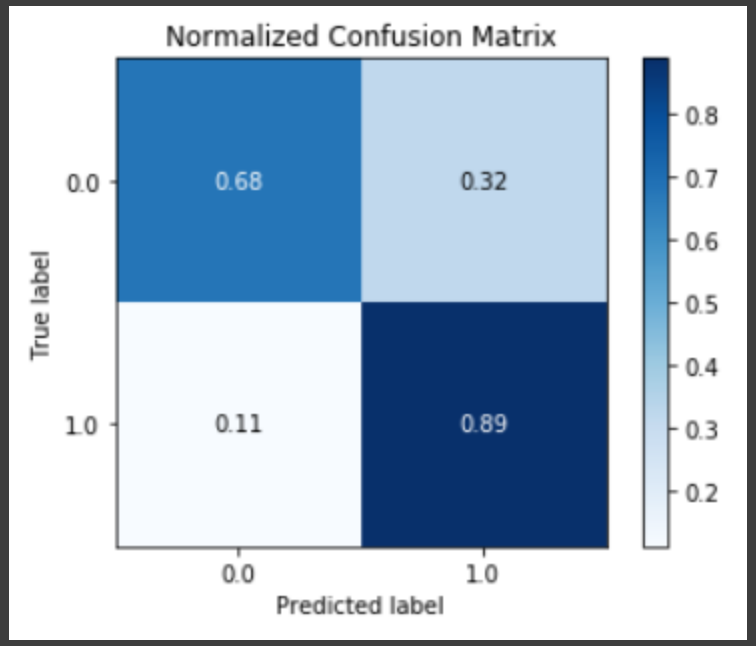


	Model	AUC
0	KNN	0.862031
1	Logistic	0.872105
2	Random_Forest	0.889422
3	SVM_Linear	0.871815
4	SVM_Kernel	0.874478

4. Modeling

Random Forest

Feature가 모두 존재하는 dataset



Confusion Matrix

Random_Forest 0.885689

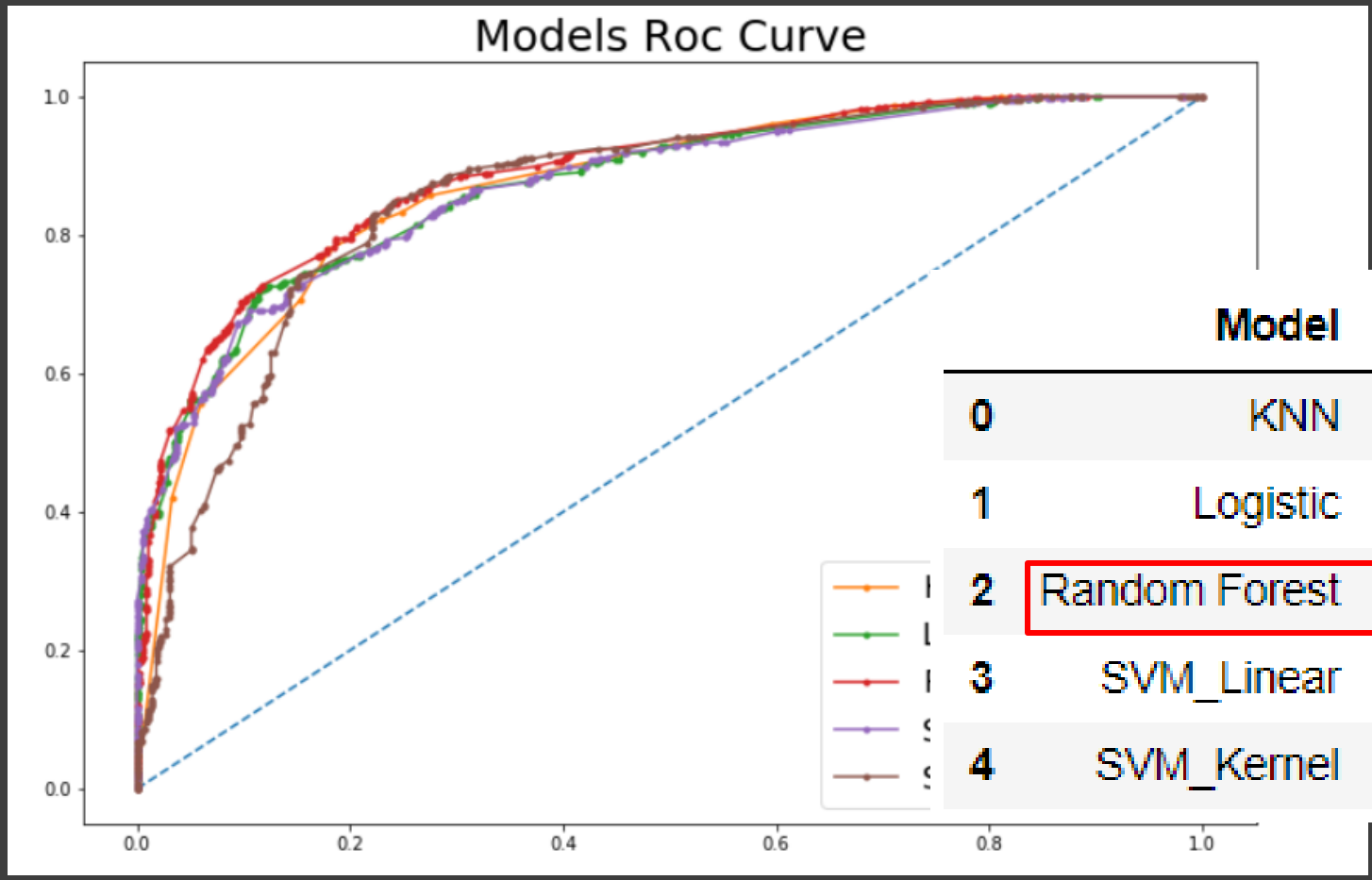
AUC 값

	precision	recall	f1-score	support
0.0	0.68	0.76	0.72	487
1.0	0.89	0.84	0.86	1080
accuracy			0.81	1567
macro avg	0.78	0.80	0.79	1567
weighted avg	0.82	0.81	0.82	1567

Classification Report

4. Modeling

Feature가 drop된 dataset



	Model	AUC
0	KNN	0.868123
1	Logistic	0.872171
2	Random Forest	0.885516
3	SVM_Linear	0.869749
4	SVM_Kernel	0.858659

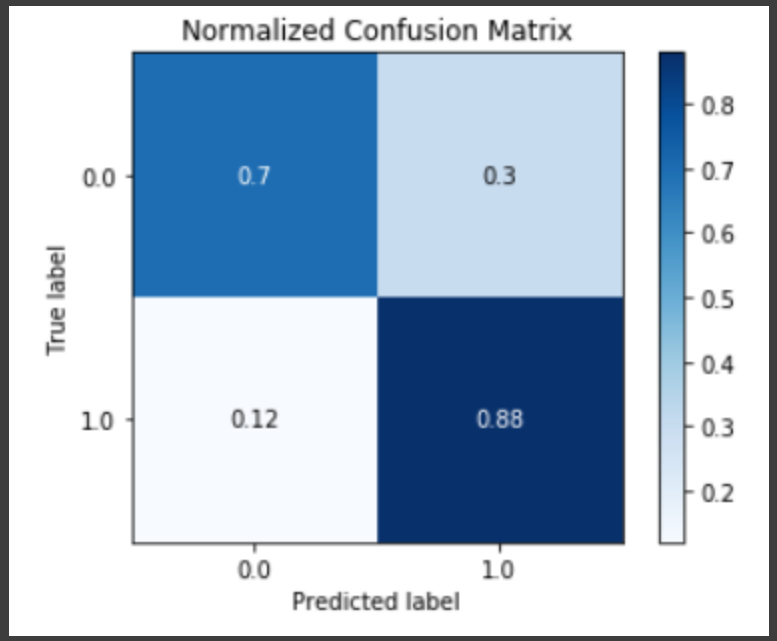
4. Modeling

Random Forest

10개의 Feature가 drop된 dataset

Random Forest 0.885516

AUC 값

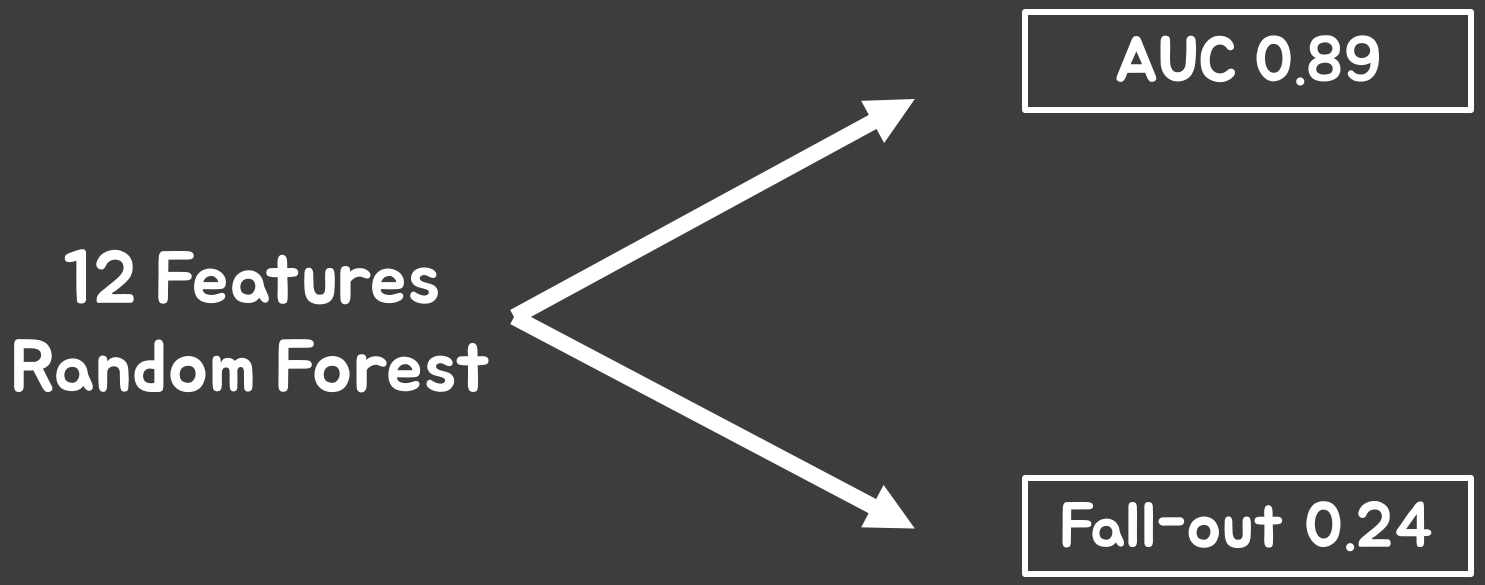


Confusion Matrix

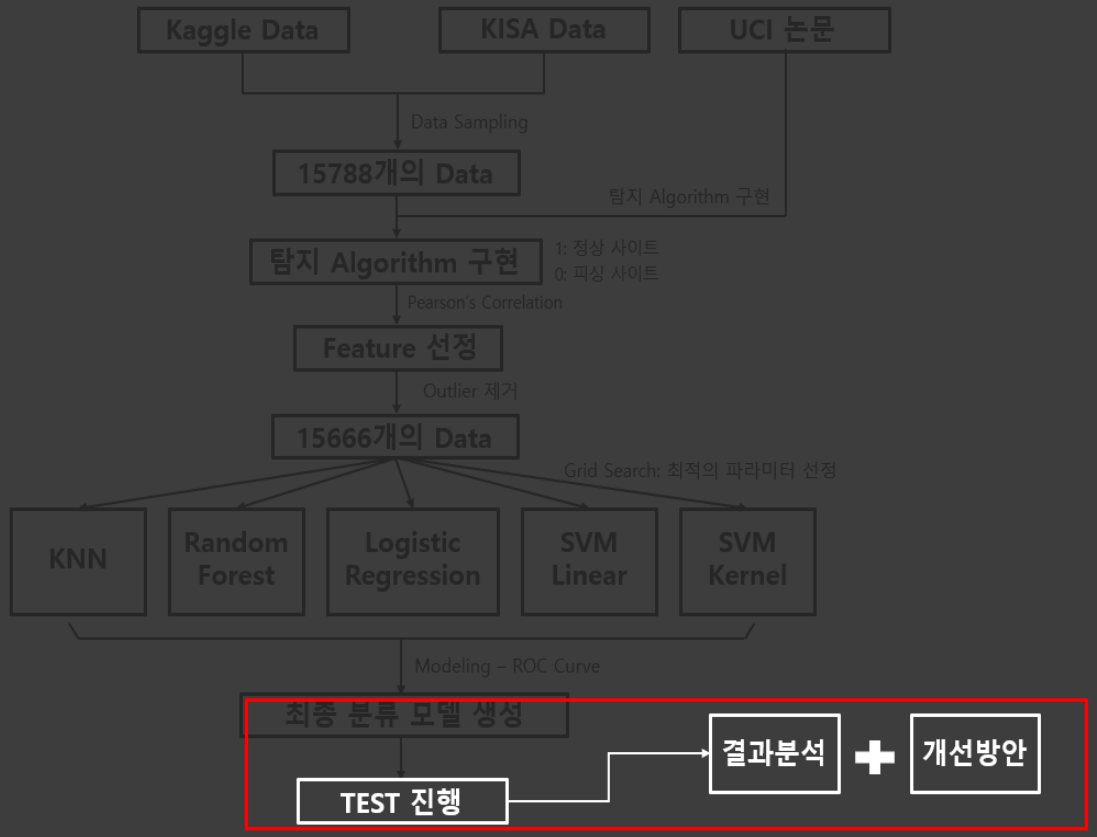
	precision	recall	f1-score	support
0.0	0.70	0.76	0.73	499
1.0	0.88	0.84	0.86	1068
accuracy			0.82	1567
macro avg	0.79	0.80	0.80	1567
weighted avg	0.82	0.82	0.82	1567

Classification Report

4. Model 성능



5. Conclusion



5-1. Test 진행

전체 Feature 사용

https://loginvkpfyn.mobile.de.a2.login.xnn-srv-as24.icu/a2/login/?
정상사이트

http://procter-gamble-session72156200.signin-openid4901473.xyz/?a=d6f636e29637c6070704e677f627261646e696c6
정상사이트

http://amazon.co.jp.5ed624a9241c819b7706b529294afb4c.icu/
피싱사이트 의심

http://rakuten.co.jp.seifioajcndsjkn.xyz/
피싱사이트 의심

http://apple.com.security-acc.email/id/?auth=HjO09r
피싱사이트 의심

http://account-update.amazon.co.jp.mtw4g-ht1-ge2rbh86rhthy.loan/
피싱사이트 의심

http://paypal.com-confirme-acount.informations-service.online/de9cc1e40c90d8cbab42b2bb0193aafaNmRjZTliMjE0YjRmNTIzYjdkNmQ4MGQwYzdmNGUzMDc=
피싱사이트 의심

---

http://www.16personalities.com/
정상사이트

http://www.nist.gov/topics/cybersecurity
정상사이트

http://mportal.ajou.ac.kr/main.do
정상사이트

http://github.com/
정상사이트

http://en.wikipedia.org/wiki/Main_Page
정상사이트

http://www.nexon.com/
정상사이트

http://kr.leagueoflegends.com/ko-kr/
정상사이트

5-1. Test 진행

10개의 Feature Drop

https://loginvkpfyn,mobile,de,a2,login,xnn-srv-as24,icu/a2/login/?
피싱사이트 의심

http://procter-gamble-session72156200,signin-openid4901473,xyz/?a=d6f636e29637c6070704e677f627261646e696c6
정상사이트

http://amazon,co,jp,5ed624a9241c819b7706b529294afb4c,icu/
피싱사이트 의심

http://rakuten,co,jp,seificajcndsjkn,xyz/
피싱사이트 의심

http://apple,com,security-acc,email/id/?auth=Hj009r
피싱사이트 의심

http://account-update,amazon,co,jp,mtw4g-ht1-ge2rbh86rhthy,loan/
피싱사이트 의심

http://paypal,com-confirme-acount,informations-service,online/de9cc1e40c90d8cbab42b2bb0193aafaNmRjZTIiMjE0YjRmNTIzYjdkNmQ4MGQwYzdmNGUzMDc=
피싱사이트 의심

---

http://www,16personalities,com/
정상사이트

http://www,nist,gov/topics/cybersecurity
정상사이트

http://mportal,ajou,ac,kr/main,do
정상사이트

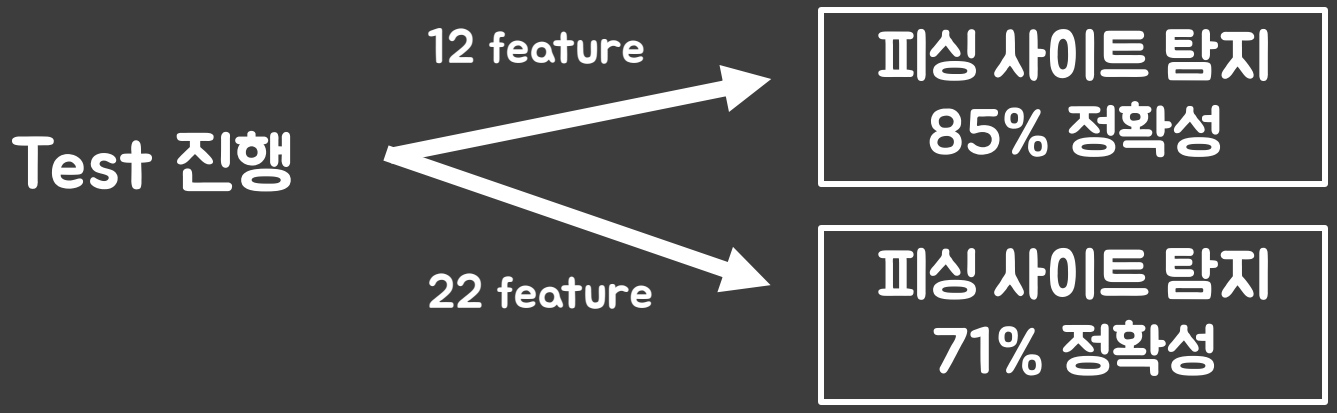
http://github,com/
정상사이트

http://en,wikipedia,org/wiki/Main_Page
정상사이트

http://www,nexon,com/
정상사이트

http://kr,leagueoflegends,com/ko-kr/
정상사이트

5-2. 결과 분석



1. Data와 feature의 전처리 과정을 통해 충분한 성능 향상을 이루어낼 수 있었음

2. 2015년에 작성된 논문을 기반으로 선정한 특징이기 때문에, 현재 생성되는 피싱 사이트 탐지에 약한 특징 존재

5-3. 개선 방안

Feature 구현
알고리즘 개선

서비스 종료된 사이트에
대한 예외처리

전처리과정 개선

Thank You