-SUPERVISED LEARNING-

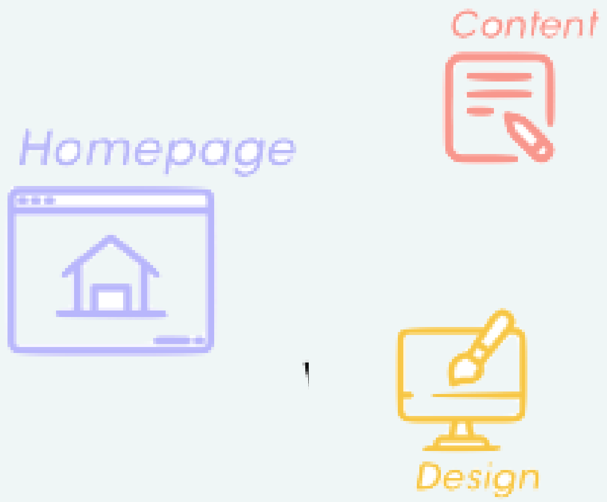
DETECTING PHISHING WEBSITES USING MACHINE LEARNING CLASSIFICATION

C-Sentinel

Your Cyber Guardian

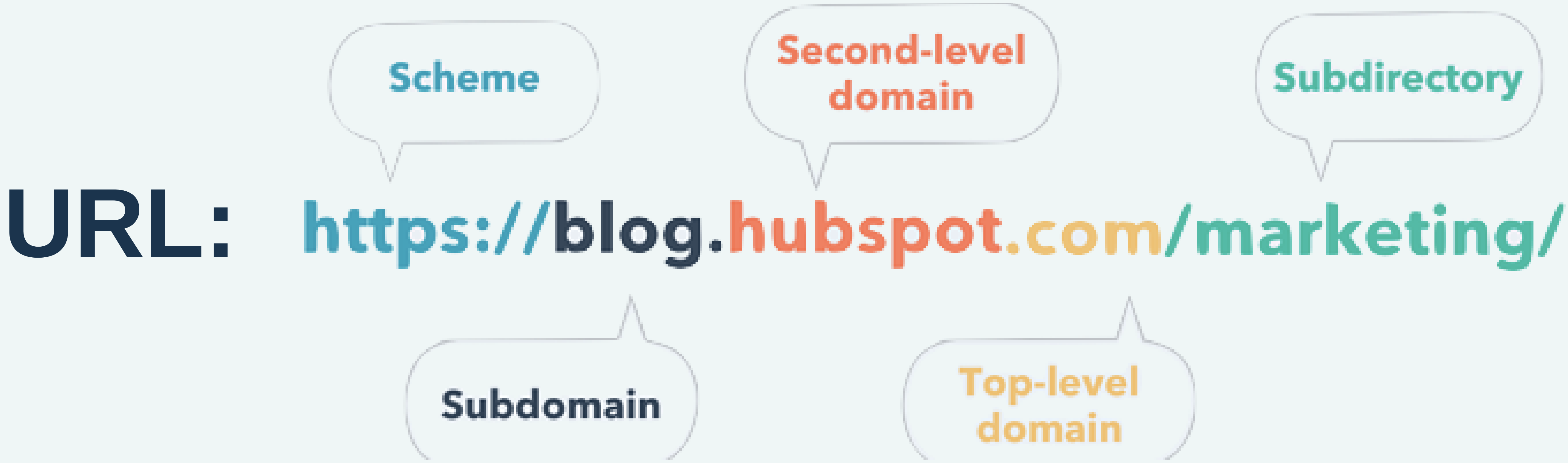
Prepared by: Haneen Alhomoud

what is a website?



WEBSITE

collection of many web pages or digital files that are written using HTML(HyperText Markup Language)


Scheme

Second-level domain

Subdirectory

URL: https://blog.hubspot.com/marketing/

Subdomain

Top-level domain

Problem Description

C-Sentinel

Cyber security company that is now looking for trustworthy and steady techniques for phishing attacks websites detection.

Phishing

Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details

Dataset

Dataset

30 Features

- UsingIP
- LongURL
- ShortURL
- Symbol@
- Redirecting//
- PrefixSuffix
- SubDomains
- HTTPS
- DomainRegLen
- Favicon
- NonStdPort
- HTTPSDomainURL
- RequestURL
- AnchorURL
- LinksInScriptTags
- ServerFormHandler
- InfoEmail
- AbnormalURL
- WebsiteForwarding
- StatusBarCust
- DisableRightClick
- UsingPopupWindow
- IframeRedirection
- AgeofDomain
- DNSRecording
- WebsiteTraffic
- PageRank
- GoogleIndex
- LinksPointingToPage
- StatsReport

Target

- Result

C-Sentinel
Your Cyber Guardian

Dataset Exploration (Data Cleansing)

Data Obtained from:

kaggle

12,000 datapoint

Data values:

1 : true
0 : false
-1 : suspicious

Outliers

None

Duplicated rows

None

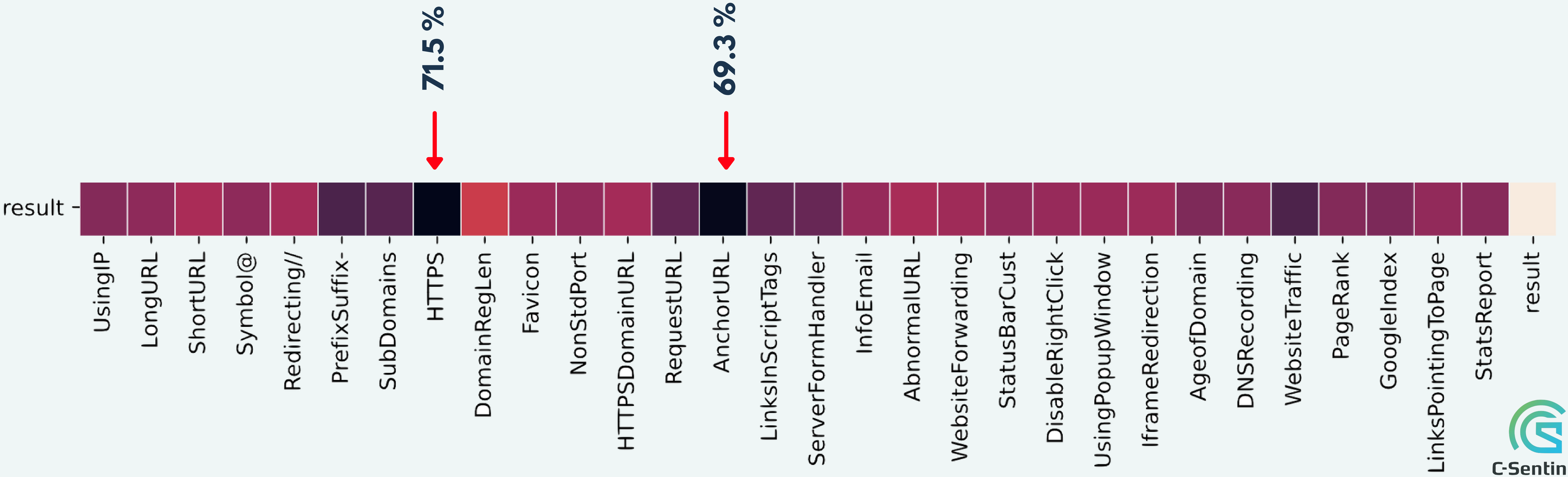
Null Values

None

C-Sentinel
Your Cyber Guardian

Dataset Correlations

- **HTTPS:** Hypertext transfer protocol secure is the secure version of HTTP, which is the primary protocol used to send data between a web browser and a website
- **AnchorURL:** An anchor is an element defined by the <a> tag



C-Sentinel
Your Cyber Guardian

Modeling - Classification

Baseline Models

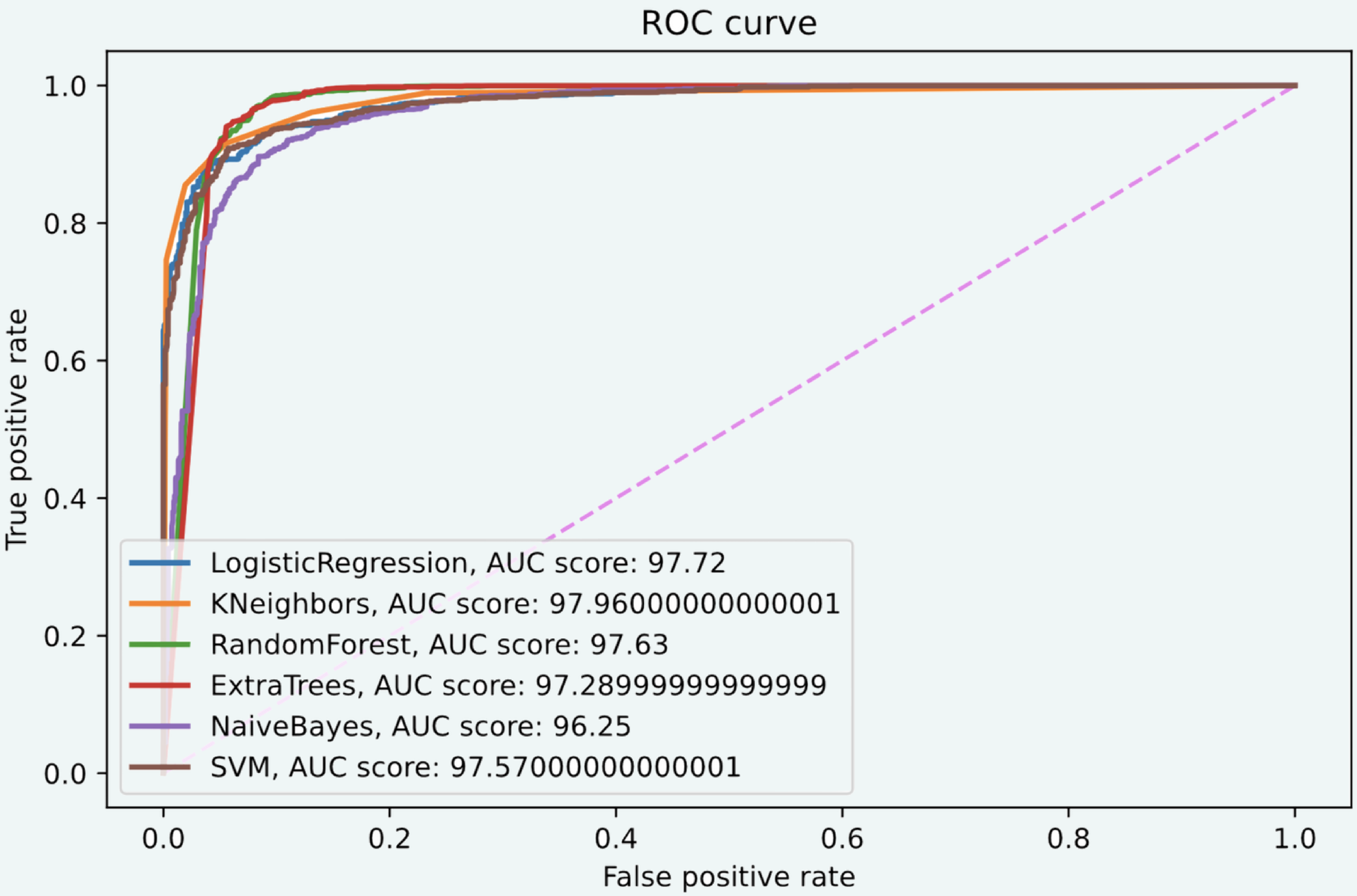
Training

	Accuracy	Precision	Recall	F2
LogisticRegression	93.2	93.4	91.1	91.6
KNeighbors	96.1	95.8	95.3	95.4
RandomForest	**99.1**	**99.1**	**99.0**	**99.0**
ExtraTrees	90.1	89.5	89.7	89.7
NaiveBayes	93.1	93.3	90.6	91.1

Validation

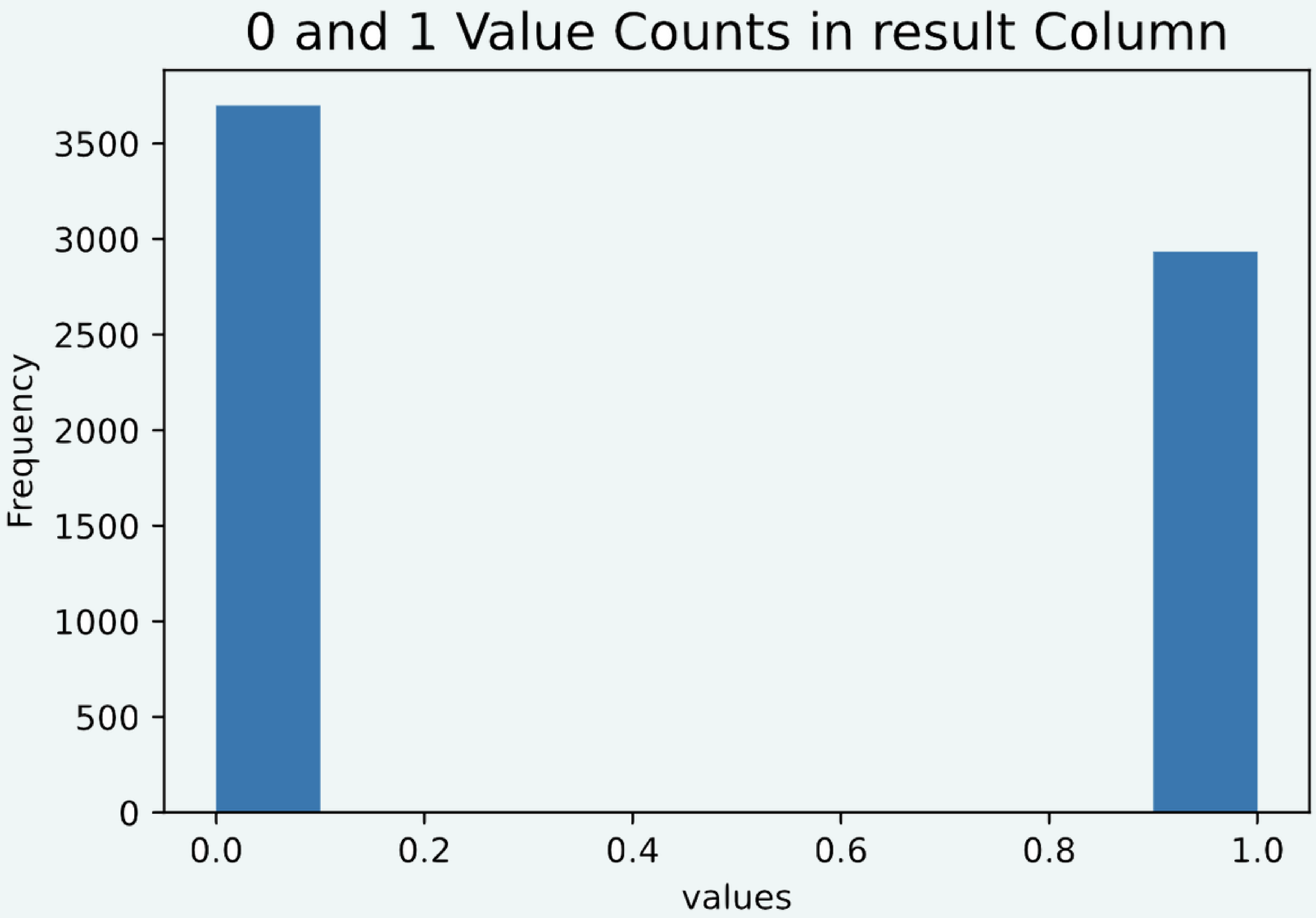
	Accuracy	Precision	Recall	F2
LogisticRegression	92.5	92.5	90.1	90.5
KNeighbors	93.4	93.1	91.6	92
RandomForest	**96.7**	**97.2**	**95**	**95.4**
ExtraTrees	96.6	97.3	94.7	95.2
NaiveBayes	92.6	93.2	89.4	90.2

Baseline Models Roc Curve



C-Sentinel
Your Cyber Guardian

Feature Engineering (Handle Imbalancement)



Legitimate websites : 3,699

Phishing websites : 2,934

C-Sentinel
Your Cyber Guardian

Upsamling

	Training f2 score	Validation f2 score
	99%	95.4%
Random	99.2	95.5
SMOTE	99.1	95.3

C-Sentinel
Your Cyber Guardian

Downsamling

	Training f2 score	Validation f2 score
	99%	**95.4%**
TomekLinks	99.1	95.5
KNN	99.9	95.7

Random Forest Hyperparameter Tuning (Random Search)

Training f2 score	99.92%	→	99.95%
Validation f2 score	95.74%	→	95.6%

C-Sentinel
Your Cyber Guardian

Conclusion and Future Work

- KNN downsamling technique gave the best score
- RandomForest Classifier gave the best score

Future Work

- Use regex to obtain all URL features
- Apply more experiment and try more models
- Autoamated API phishing web detection

C-Sentinel
Your Cyber Guardian

Thank You