Analysis of normal and phishing URLs

Group 4 : Hyun woo Jung, Jung min Lee, Jin young Kim

Date : 2025.07.11

UtahState University

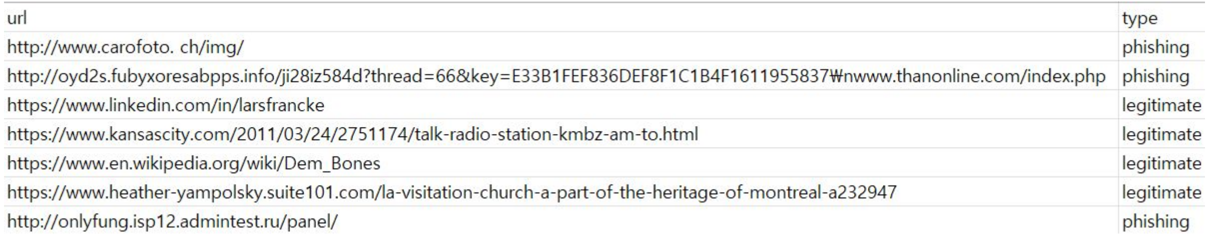
Data Description

Data source : KAITHOLIKKAL, JISHNU K S; B, Arthi (2024), “Phishing URL dataset”, Mendeley Data, V1, doi: 10.17632/vfszbj9b36.1

Total number of samples : 2,008,874

Class distribution : with 76.8% legitimate and 23.2% phishing URLs. After balancing, the dataset contains a 50:50 ratio.

Data Preprocessing : Missing value removal, Class balancing(5:5 ratio), Feature extraction from long URLs.



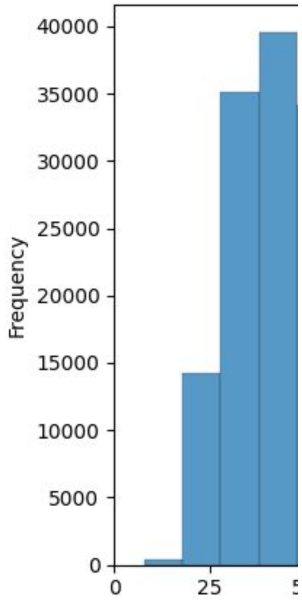
url	type
http://www.carofoto. ch/img/	phishing
http://oyd2s.fubyxoresabpps.info/ji28iz584d?thread=66&key=E33B1FEF836DEF8F1C1B4F1611955837₩nwww.thanonline.com/index.php	phishing
https://www.linkedin.com/in/larsfrancke	legitimate
https://www.kansascity.com/2011/03/24/2751174/talk-radio-station-kmbz-am-to.html	legitimate
https://www.en.wikipedia.org/wiki/Dem_Bones	legitimate
https://www.heather-yampolsky.suite101.com/la-visitation-church-a-part-of-the-heritage-of-montreal-a232947	legitimate
http://onlyfung.isp12.admintest.ru/panel/	phishing

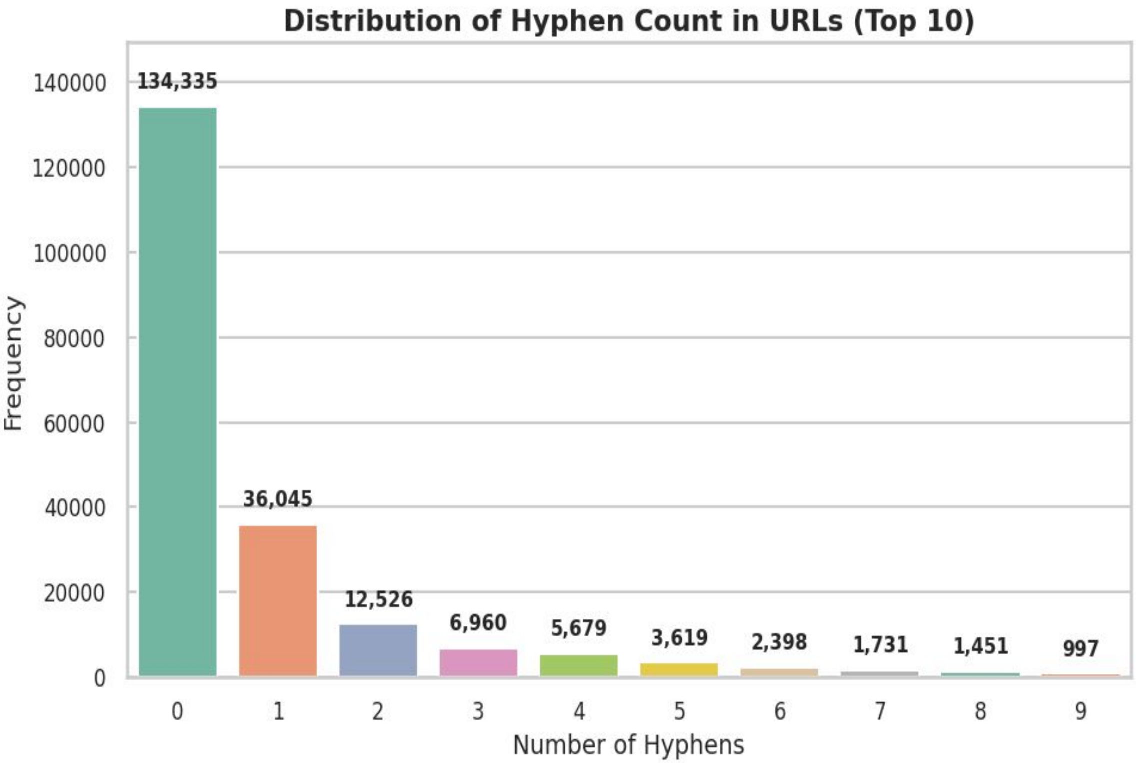
Preprocessing - Feature Extraction

Feature	Example	meaning
url_length	30	Total number of characters in the URL
num_dots	3	Number of dots ‘.’ in the URL
num_hyphens	1 or 2	Number of dots ‘-’ in the URL
has_https	False(0)	Whether the URL starts with ‘ https:// ’
has_suspicitous_word	True(1)	Value: 1 if suspicious word is present, 0 otherwise Keywords: 'secure', 'account', 'update', 'free', 'bonus', 'verify', 'signin', 'bank', 'confirm'
type	legitimate phishing	normal URLs dataset and phishing URLs dataset

Feature - Visual Analysis

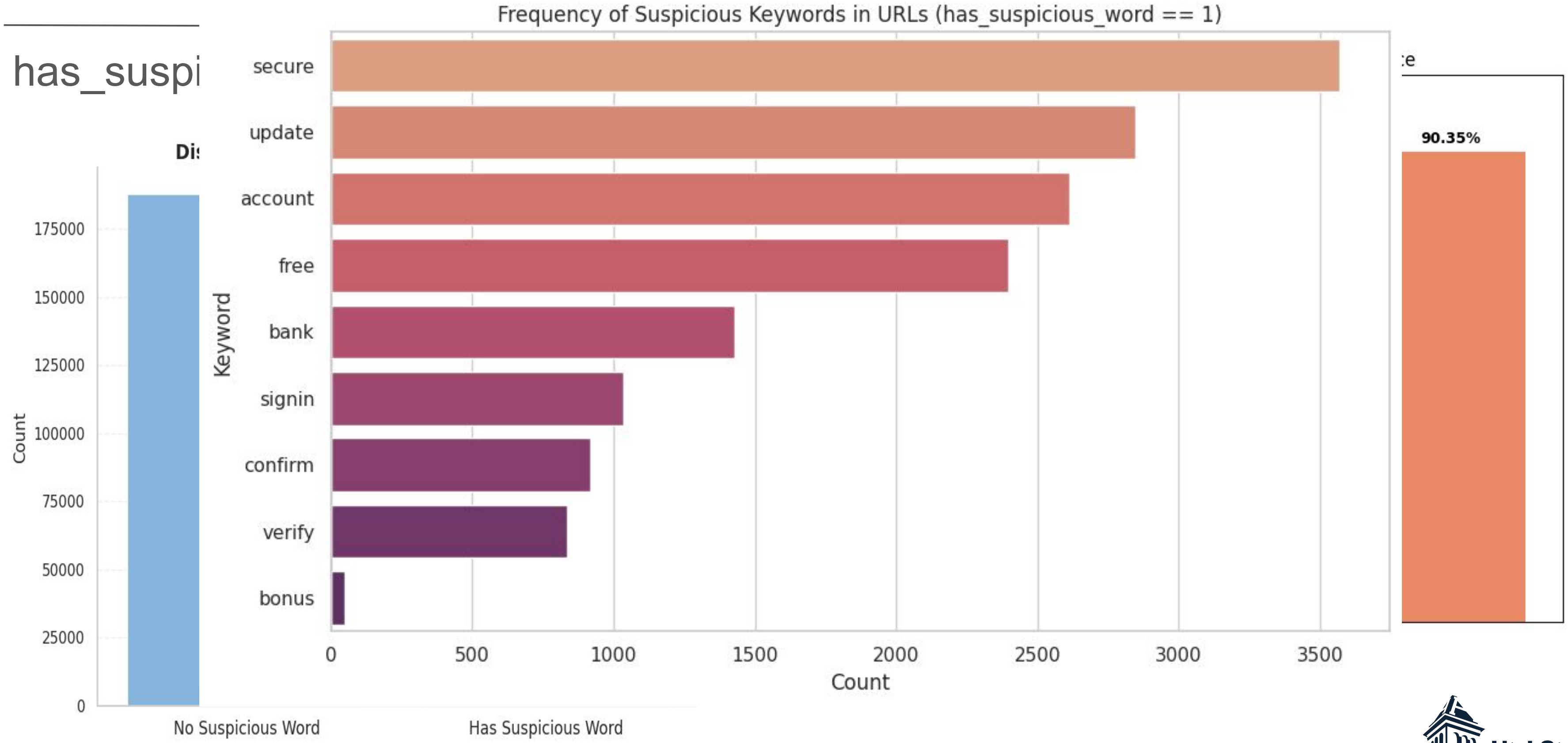
url_length





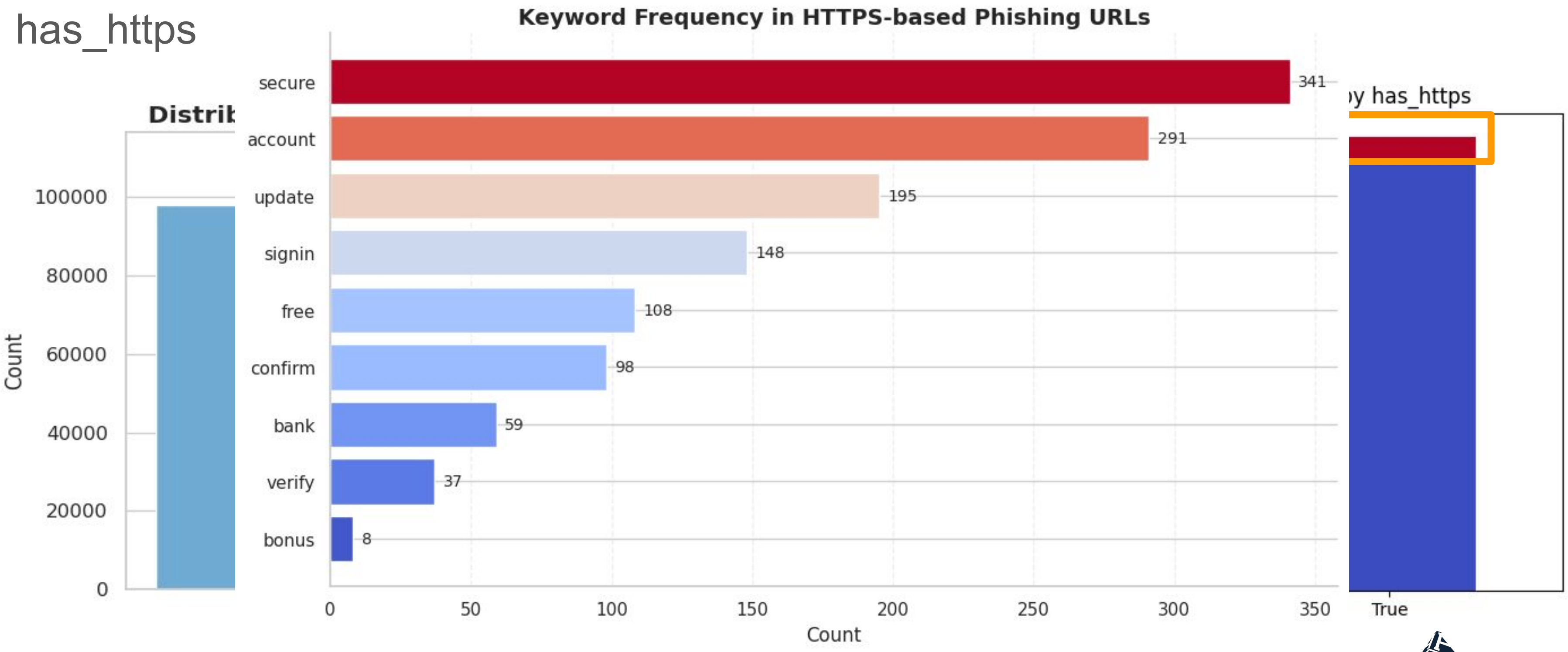
Feature - Visual Analysis

has_suspi

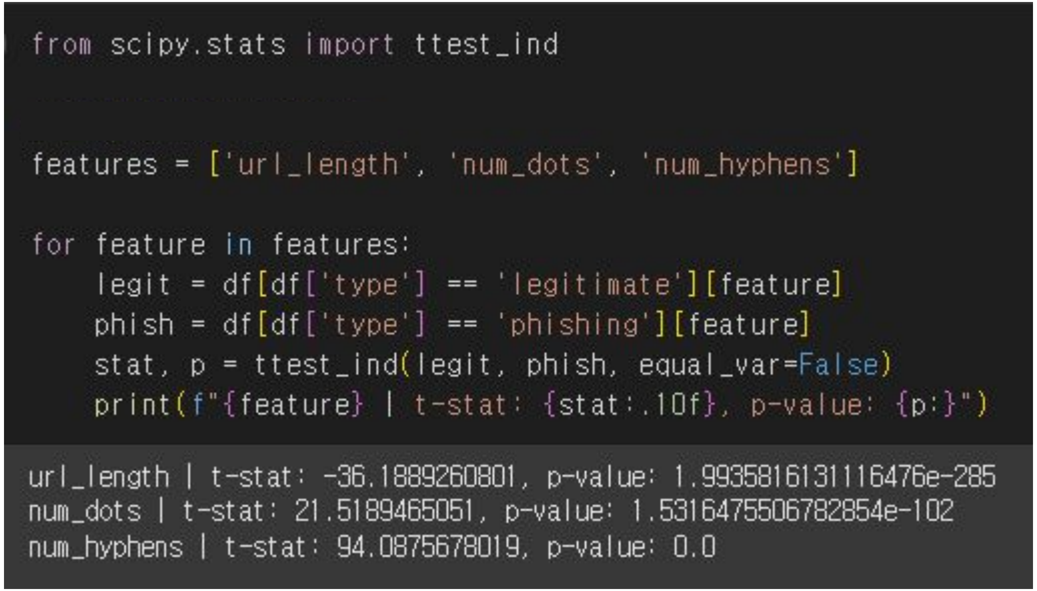


Feature - Visual Analysis

has_https



Result - Using T-test



```
from scipy.stats import ttest_ind


features = ['url_length', 'num_dots', 'num_hyphens']

for feature in features:
    legit = df[df['type'] == 'legitimate'][feature]
    phish = df[df['type'] == 'phishing'][feature]
    stat, p = ttest_ind(legit, phish, equal_var=False)
    print(f"{feature} | t-stat: {stat:.10f}, p-value: {p:}")
```

```
url_length | t-stat: -36.1889260801, p-value: 1.9935816131116476e-285
num_dots | t-stat: 21.5189465051, p-value: 1.5316475506782854e-102
num_hyphens | t-stat: 94.0875678019, p-value: 0.0
```

Performed t-tests for each feature

All features showed **very small p-values (p < 0.001)**

➤ Indicates **strong difference** between phishing and legitimate URLs

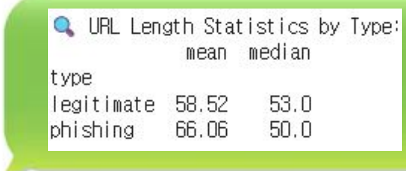
🔍 All features were **statistically significant → kept for modeling**

Result - Correlation Coefficient

```
corr = df[['url_length', 'num_dots', 'num_hyphens', 'has_https', 'has_suspicious_word',
'type_binary']].corr()
print(corr['type_binary'].sort_values(ascending=False))
```

```
🔍 Correlation with type (phishing=1):

has_suspicious_word     0.267906

url_length              0.078937

num_dots               -0.047033

num_hyphens            -0.201641

has_https              -0.939625

Name: type_binary, dtype: float64
```

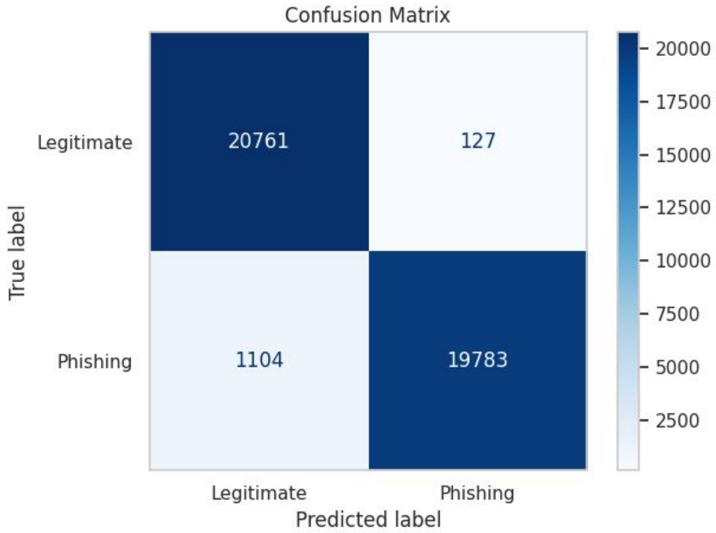
Most individual features show **weak correlation with phishing labels** (e.g., `url_length` = 0.08, `has_suspicious_word` = 0.27).
This indicates that a multivariate approach, such as logistic regression, **is necessary for more accurate classification.**

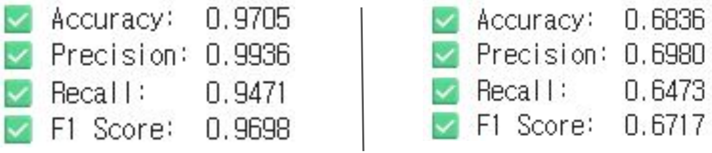


```
🔍 URL Length Statistics by Type:
             mean  median
type
legitimate  58.52    53.0
phishing    66.06    50.0
```

Additional Results - Logistic Regression

```
features = ["url_length", "num_dots",
"num_hyphens", "has_https",
"has_suspicious_word"]
X = df[features]
y = df["type_binary"]

X_train, X_test, y_train, y_test =
train_test_split(
    X, y, test_size=0.2, random_state=42,
stratify=y
)

model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```





```
✅ Accuracy:  0.9705
✅ Precision: 0.9936
✅ Recall:    0.9471
✅ F1 Score:  0.9698
```

```
✅ Accuracy:  0.6836
✅ Precision: 0.6980
✅ Recall:    0.6473
✅ F1 Score:  0.6717
```

Conclusion

Caution on HTTPS

While HTTPS presence generally indicates a legitimate site, our analysis shows some phishing URLs still use HTTPS — so HTTPS alone is not a reliable indicator.

Statistical Significance:
T-tests confirmed that url_length, num_dots, and num_hyphens show statistically significant differences between phishing and legitimate URLs.

Correlation Insight:
Correlation analysis revealed weak individual feature associations (e.g., url_length = 0.08, has_suspicious_word = 0.27), indicating the need for a multivariate approach.

Model Performance:
Logistic regression achieved high performance (Accuracy: **97.05%**, F1 Score: **96.98%**), proving that combining features enhances phishing detection.

Thank you.

UtahState
University