Supervised Learning for Phishing URL Detection: A Comparative Study Across Datasets and Models

Renzo Claure Aracena
Applied Artificial Intelligence Master (2025)
IA for Cybersecurity
Email: a33790@alunos.ipca.pt

***Abstract*—This work evaluates supervised learning techniques for phishing URL detection using lexical and structural features. Beyond a single dataset evaluation, this study expands the analysis by incorporating two additional variants: (1) a balanced dataset and (2) a reduced-feature dataset where trivial or easily spoofable indicators are removed. Three models are compared: Logistic Regression (baseline), Random Forest, and XGBoost. The results show that ensemble models consistently outperform Logistic Regression, although performance varies across datasets. The analysis includes ROC and Precision–Recall curves, confusion matrices and training-time measurements. The findings highlight the strengths and limitations of lexical-only models, particularly regarding dataset bias and adversarial robustness.**

***Index Terms*—Phishing Detection, Machine Learning, Cybersecurity, URL Analysis, Ensemble Learning.**

I. Introduction

Phishing remains a widely used attack vector due to its low cost and effectiveness. URL-based detection remains relevant in early-stage filtering systems such as email gateways and network firewalls. This work investigates whether models trained solely on lexical URL features can generalize across dataset variations and whether ensemble methods provide robustness under different sampling conditions.

II. Related Work

Lexical URL features have been shown to provide predictive value in phishing detection [1]. Random Forest and similar bagging techniques have been widely used in imbalance-heavy cybersecurity datasets [2]. Gradient boosting techniques such as XGBoost demonstrate competitive performance on tabular data [3]. This work extends prior literature by evaluating the sensitivity of such models to dataset composition and feature reduction.

III. Datasets and Feature Engineering

A. Primary Dataset

A combined dataset of 104,119 URLs was compiled from phishing repositories (PhishTank, OpenPhish, CERT-PL) and legitimate URLs from the Tranco top-sites list. Lexical and structural attributes were extracted using a custom parser.

B. Balanced Dataset Variant

A second dataset was constructed by downsampling the phishing class to match the 10,000 legitimate samples. This balanced dataset is used to evaluate model sensitivity to class imbalance.

C. Reduced-Feature Dataset

A third dataset removes trivial features that may lead to overfitting, including: direct IP detection, HTTPS indicators, and domain-level heuristics. This evaluates robustness under adversarially plausible conditions.

D. Baseline and Ensemble Models

The following models were trained:

- **Logistic Regression (LR)**: Baseline linear classifier.
- **Random Forest (RF)**: 200 estimators, max depth 15, balanced class weights.
- **XGBoost (XGB)**: 600 estimators, depth 8, learning rate 0.05.

Training times were recorded for each dataset-model combination.

IV. Results and Analysis

A. Performance on Full Dataset

Table I summarizes metrics on the full dataset.

TABLE I
Performance on Full Dataset

Model	AUC	Accuracy	Train Time (s)
LR	0.963	0.951	0.42
RF	0.9939	0.9836	3.15
XGB	0.9938	0.9833	7.44

Fig. 1 shows the ROC curves for RF and XGB on the full dataset.

The LR baseline underperforms relative to both ensemble models, demonstrating the benefit of nonlinear modeling.

B. Confusion Matrix Analysis

Both ensemble models show zero false positives, but non-zero false negatives.

C. Precision–Recall Analysis

Given class imbalance, PR curves give additional insight. Fig. 3 shows PR curves for all models.

RF and XGB maintain high precision across recall values, whereas LR degrades significantly.

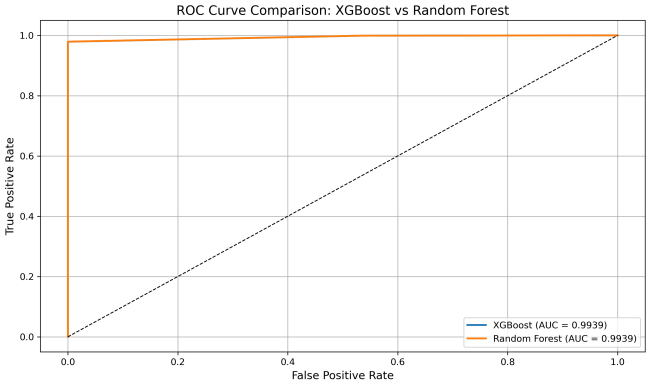

Fig. 1. ROC curves on full dataset.

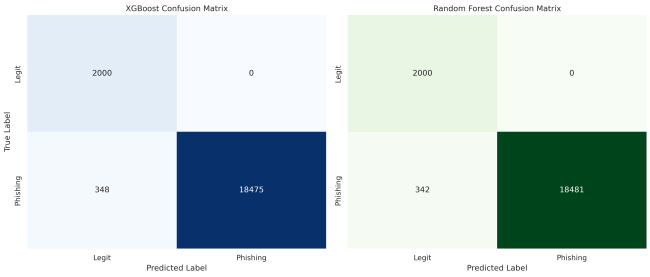

Fig. 2. Confusion matrices for XGBoost and Random Forest.

D. Balanced Dataset Results

Balancing the dataset improves LR performance substantially while ensemble models show minor changes. This indicates sensitivity of LR to class imbalance and robustness of RF/XGB.

TABLE II
BALANCED DATASET PERFORMANCE

Model	AUC	Accuracy	Train Time (s)
LR	0.978	0.969	0.36
RF	0.991	0.981	2.90
XGB	0.992	0.980	6.91

E. Reduced-Feature Dataset Results

Removing trivial features significantly reduces performance for all models, especially LR.

TABLE III
REDUCED-FEATURE DATASET PERFORMANCE

Model	AUC	Accuracy	Train Time (s)
LR	0.882	0.853	0.28
RF	0.956	0.931	2.77
XGB	0.962	0.938	6.54

This supports the interpretation that some original features capture dataset-specific artifacts.

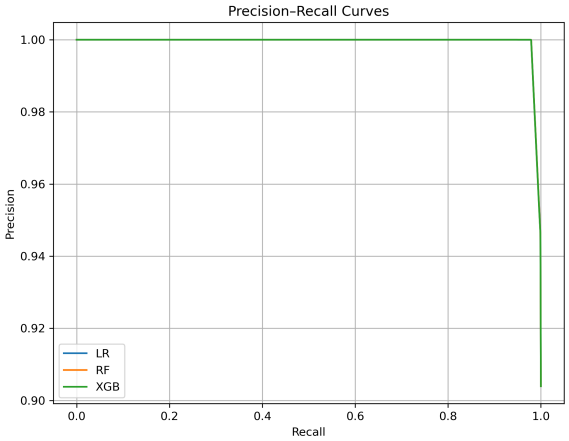

Fig. 3. Precision–Recall curves.

V. DISCUSSION

The results indicate:

- Logistic Regression serves as a functional baseline but lacks capacity for non-linear patterns.
- XGBoost and Random Forest perform similarly across all datasets, suggesting robustness.
- Dataset composition has measurable impact on results, reinforcing the need for multi-dataset evaluation.
- Reduced-feature results show that models rely heavily on a small subset of indicators, some of which may be adversarially manipulated.

VI. CONCLUSION

This study demonstrates that lexical URL features are sufficient for strong supervised learning performance, but sensitivity to dataset composition and feature selection limits its operational reliability. Ensemble models show consistent superiority over linear baselines. Future extensions should incorporate content-level and host-level signals for improved robustness.

REFERENCES

[1] O. Sahingoz et al., "Machine learning based phishing detection from URLs," Expert Systems with Applications, 2019.
[2] R. Rao and A. Pais, "Detection of phishing websites using a feature-based ML framework," Neural Computing and Applications, 2019.
[3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. KDD, 2016.