Think Before You Click: A Machine Learning Approach to Phishing URL Detection

Navya Shiju
Dr. Mark Zarb

Introduction

The detection of phishing URLs has become an important area of research within the cyber-security community, as traditional methods for detecting phishing websites have become increasingly ineffective in combating these evolving threats.

This project aims to understand how a hybrid model performs in terms of real-time efficiency when compared to individually trained models, and to also understand which models will have the highest accuracy in predicting the authenticity of a Phishing URL?

Data Pre-Processing
Relevant URL Feature Extraction and Standardisation

Baseline Model Implementation
Decision Tree, Random Forest, SVM

Model Optimisation
Validation Method Incorporation:
Stratified K-Fold Cross-Validation (10 Splits)

Model Optimisation
Hyper-Parameter Tuning - GridSearchCV

Hybrid Model Implementation
Optimised SVM + Random Forest -> XGBoost Classifier

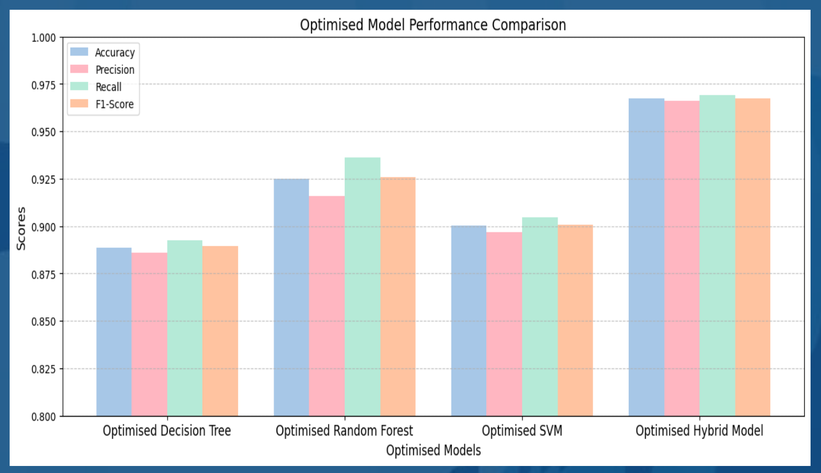
Evaluation
Accuracy, Precision, Recall, F1-Score, Inference Time

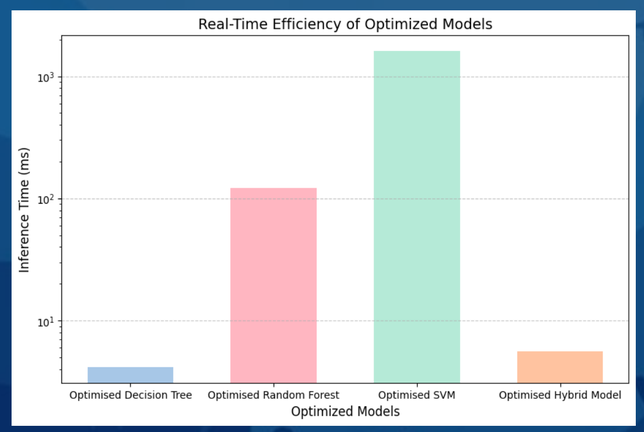
Methodology

The chosen methodology follows a structured pipeline comprising of data pre-processing, baseline model development and evaluation, cross-validation, hyper-parameter tuning, hybrid model implementation, and final model evaluation.

Results

The hybrid model outperformed all individual models, achieving 96.76% accuracy. This confirms that combining Random Forest and SVM, with XGBoost as the meta-classifier, results in superior predictive performance than individually trained models. However, with 92.51% accuracy, Random Forest proved to be the most effective standalone classifier.





The Optimized Hybrid Model is the most suitable choice for corporate cybersecurity applications, as it delivers high detection accuracy with minimal inference time.

The findings confirm that ensemble learning and hybrid stacking significantly improve detection accuracy.

Future Work

Future work will explore deep learning models to improve phishing URL detection by capturing sequential dependencies and contextual patterns within URLs, potentially enhancing classification accuracy against more sophisticated attacks.

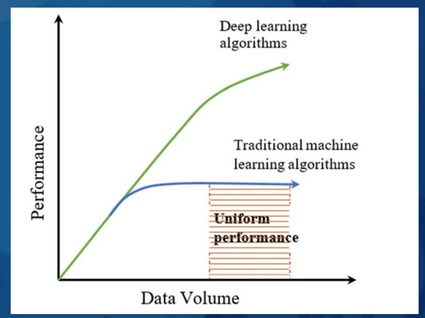


Image sourced from Ahmed, S.F. et al. (2023)

The implementation of a browser extension within corporate networks will be considered to provide proactive phishing prevention by analysing and blocking suspicious URLs in real time.

Acknowledgments

I would like to extend my gratitude to my supervisor and The Zarbers for their ongoing support, and unwavering optimism in the face of all my panicked babble.

SCHOOL OF COMPUTING, ENGINEERING & TECHNOLOGY

Ahmed, S.F. et al. (2023) "Deep learning modelling techniques: current progress, applications, advantages, and challenges," Artificial intelligence review, 56(11), pp. 13521–13617. Available at: https://doi.org/10.1007/s10462-023-10466-8.

MSci Computing Science