PHISHGUARD: A MULTI-LAYERED ENSEMBLE MODEL FOR OPTIMAL PHISHING WEBSITE DETECTION

Md Sultanul Islam Ovi
movi@gmu.edu

CS 700
Fall 2023

Introduction

- Escalating sophistication and frequency in cyber attacks
- Financial and data losses from phishing incidents
- Traditional methods inadequate against new phishing tactics
- Multi-layered ensemble model for accurate, efficient detection

Techniques and System Analysis

- **Dataset**
  - Phishing website dataset *
    (11055 x 32)

- **Machine Learning Models**
  - SVM
  - Random Forest
  - XGBoost

- **Data Preprocessing**
  - Data Sampling
  - Feature Selection
  - Mean, Median, Quartiles - Box Plots

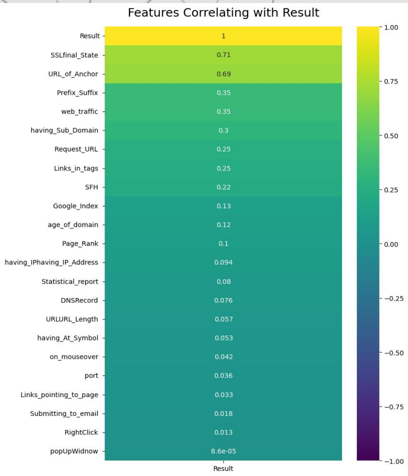
- **Performance Analysis**
  - Hypothesis Testing
  - Confidence Intervals
  - ANOVA Test
  - Contrast Approach

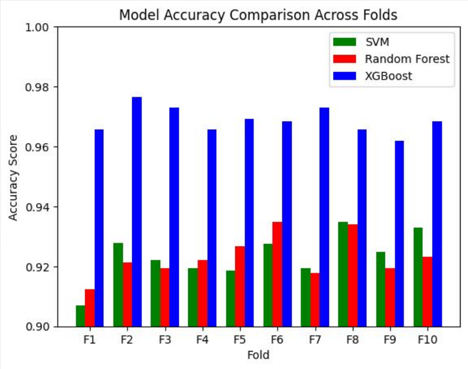
★ https://www.kaggle.com/datasets/akashkr/phishing-website-dataset/data

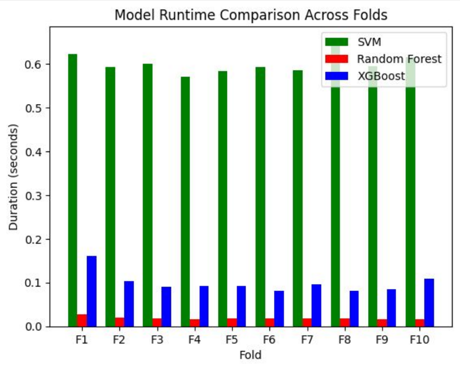
Methodological Approach

- ❑ Balanced dataset sampling followed by feature selection using Recursive Feature Elimination (RFE).
- ❑ Utilized box plots to represent data distribution, outliers, and feature relationships.
- ❑ Trained SVM, Random Forest, and XGBoost models using 10-fold cross-validation, focusing on accuracy and training/testing duration.
- ❑ Utilized hyperparameter tuning (grid search) to optimize model performance, alongside metrics like precision, recall, and F1-score.
- ❑ Conducted hypothesis testing and calculated confidence intervals.
- ❑ Applied ANOVA and contrast approach to evaluate and compare model performances. (significance level = 5%)

Preliminary Results







F-Statistic	154.15584
P-Value	1.7e-15
Decision	Significant differences
CI of SVM	(0.91769, 0.92927)
CI of Random Forest	(0.91808, 0.92814)
CI of XGBoost	(0.96560, 0.97180)

Table: ANOVA Test

Comparison	**Confidence Interval**	**Decision**
SVM vs Random Forest	(-0.00441, 0.00514)	No significant difference
Random Forest vs XGBoost	(-0.05037, -0.04081)	XGBoost better
SVM vs XGBoost	(-0.05000, -0.04045)	XGBoost better

Table: Contrast Approach