Report: Insights on Phishing Website Detection Dataset

1. Data Overview

- **Dataset Source**: UCI Machine Learning Repository – Phishing Websites Dataset (ID: 327).
- **Shape**: Initially includes 11,055 rows and 31 features.
- **Target Variable**: Binary classification — `1` for Phishing and `-1` for Legitimate.
- **Data Cleaning**:
  - Missing values: None found.
  - Duplicate entries: Identified and removed.
  - Data was exported as `'Phishing Websites Preprocessed.csv'`.

---

2. Feature Engineering

Created meaningful aggregate and scoring features to improve interpretability and model performance:

1. **`total_link_flags`**
   Sum of link-related features like `request_url`, `url_of_anchor`, `links_in_tags`, `statistical_report`.
2. **`security_score`**
   Average score from security-related indicators such as `sslfinal_state`, `https_token`, `dnsrecord`, etc.
3. **`obfuscation_score`**
   Captures the level of URL manipulation (e.g., `having_at_symbol`, `prefix_suffix`, `url_length`).
4. **`tech_complexity`**
   Combines features reflecting web page complexity and behavior: `sfh`, `iframe`,

`rightclick`, etc.

Insight: These engineered features enhance model explainability by grouping semantically related features into interpretable categories.

---

3. Feature Selection Techniques for Explainability

Several statistical techniques were applied to rank feature importance and evaluate their relevance to the target:

---

a. Variance Threshold

- Removed features with near-zero variance (i.e., no discriminatory power).
- **Visualization**: Top 10 features plotted by variance.

Insight: Helped eliminate redundant features and focus on those with more variability.

---

b. Chi-square Test

- Assesses dependency between features and the (binarized) target.
- **Findings**: High scores suggest strong association with phishing or legitimate class.
- **Visualization**: Bar chart of chi-square scores.

Insight: Provided an interpretable ranking of categorical features in terms of their discriminatory power.

---

c. ANOVA F-test

- Evaluates differences in feature means across the two classes.
- **Visualization**: Bar chart showing F-scores for all features.

Insight: Features with higher F-scores are more statistically significant in distinguishing classes.

---

d. Mutual Information (MI)

- Measures mutual dependence between features and the target.
- **Top Features (MI)**: Listed in a bar plot.

Insight: MI helped uncover non-linear dependencies often missed by correlation-based methods.

---

e. Fisher's Score

- Ratio of inter-class variance to intra-class variance.
- **Visualization**: Fisher scores plotted for interpretability.

Insight: Strong discriminators show high Fisher scores; particularly helpful in binary classification.

---

f. Correlation with Target

- Standard Pearson correlation between features and the target class.
- **Visualization**: Bar plot of correlation coefficients.

Insight: Quickly highlights linear relationships. Complementary to MI.