Extended Explainability Report: Tree-Based Model – XGBoost Classifier

Model Summary

Model	Description
XGBoost Classifier (XGB)	Gradient-boosted decision tree ensemble — balances bias-variance, handles feature interactions well, and often outperforms single-tree models.

Model Performance

Metric	Value
Accuracy	96%

Comments:

- High accuracy with good generalization.
- Robust against overfitting due to boosting with regularization.

Explainability with SHAP

XGBoost SHAP Summary

- **SHAP explainer** used on `.predict_proba()` for better class-level insights.
- **Top Influencers** (Global Feature Importance):
    - `URL_of_Anchor`

    - `Prefix_Suffix`
    - `Request_URL`
    - `SFH`
    - `web_traffic`

Insights:

- Feature effects are **directional**: e.g., high values of `Prefix_Suffix` push toward phishing class.
- SHAP plots are **highly interpretable**, showing which features increase or decrease phishing likelihood.
- SHAP confirmed consistent importance across samples.

Explainability with LIME

XGBoost (LIME)

- Local explanations generated on **random test instance**.
- **Top Local Influencers** (Instance-level):
    - `Request_URL`
    - `having_Sub_Domain`
    - `URL_of_Anchor`
    - `web_traffic`

Insights:

- LIME matched SHAP in identifying critical features.
- Clear visualization of **how specific feature values** impacted the predicted class probability.
- Effective for **per-instance storytelling** (why *this* prediction was made).

Permutation Feature Importance (PFI)

- Measures drop in performance when a feature is permuted (shuffled).
- **Top Features by Importance**:
  1. `URL_of_Anchor`
  2. `SFH`
  3. `Prefix_Suffix`
  4. `web_traffic`
  5. `Request_URL`

Insights:

- Confirms SHAP's and LIME’s findings.
- `URL_of_Anchor` consistently impacts model performance.

Leave-One-Feature-Out (LOFO) Importance

- Evaluates performance drop when one feature is removed at a time.
- **Top Features by LOFO Impact**:

1. `Prefix_Suffix`
2. `Request_URL`
3. `URL_of_Anchor`
4. `having_Sub_Domain`
5. `web_traffic`

Insights:

- Removing these features notably reduced cross-validation accuracy.