Model Performance & Explainability Report (Stacking Classifier)

Dataset Overview

- **Dataset**: Phishing Websites (Engineered)
- **Target**: Binary classification Phishing (1) vs. Legitimate (-1)
- **Train-Test Split**: 80% training, 20% testing

Model Used

- **Classifier**: Stacking Classifier
    - **Base Models**: Random Forest, Gradient Boosting
    - **Final Estimator**: Logistic Regression
- **Rationale**: Combining multiple strong learners also mixes and multiplies their strengths. The models further benefit from generalization improvements.

Performance Metrics (Test Set)

Metric	Value
Accuracy	0.9393
Precision	0.9412
Recall	0.9329
F1 Score	0.9370

- Stacking classifiers tend to work best and sometimes slightly outperform individual models in F1 score and balance.

Confusion Matrix

- This shows that performance in terms of precision and recall is quite high.

- It has shown a low misclassification rate on both legitimate and phishing classes.

Explainable AI (XAI) Techniques

1. SHAP (Kernel Explainer)

- **Type**: Model-agnostic (Kernel SHAP)
- Applied to a sample of 10 instances
- **Insights**:
    - SHAP values show which features mostly affect the prediction of the phishing class.
    - Effects of features are visualized by their magnitude and direction.
    - Example: An increase in URL_of_Anchor values may increase phishing susceptibility.

---

2. LIME (Local Interpretable Model-Agnostic Explanations)

- **Type**: Local explanation
- Applied to one random test instance
- **Output**:
    - The LIME exhibits the top 10 features that contributed to the predicted category.

- It helps you to understand how the stack model reached a decision.

---

3. Permutation Feature Importance (PFI)

- **Type**: Global, model-agnostic
- Evaluates the decrease in performance when feature values are shuffled
- **Top 5 Features**:
    1. URL_of_Anchor
    2. SSLfinal_State
    3. Prefix_Suffix
    4. Web_Suffix
    5. Total_Link_Flags

Interpretation: These features are most critical for the model's predictive performance.

---

4. Partial Dependence Plots (PDP) & ICE

- **Features Analyzed**: Top 5 by PFI, filtered to continuous-valued features
- **Insights**:
    - PDP line shows average model output as a function of a single feature
    - ICE lines show how individual predictions change, offering a personalized view

---

Conclusion

The Stacking Classifier Model showed very good performance in detecting potential phishing cases. The strength of its explainability comes from the fact that all insights emanating from SHAP, LIME, PDP or ICE, and PFI were conflicting considering the highest contributing features across-all conditions were aligned, thus increasing the fear that has been built around the transparency of the model and its applicability for cybersecurity.