Model Performance & Explainability Report

Dataset Overview

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

---

Model Used

- **Classifier**: Gradient Boosting Classifier
- **Rationale**:As for tabular data, many times, it gives a good performance in classification tasks.

Performance Metrics (Test Set)

Metric	Value
Accuracy	0.9402
Precision	0.9291
Recall	0.9488
F1 Score	0.9388

- Precision and recall being high could mean that there was less chance of false positives or negatives.

Confusion Matrix

- Clear separation between legitimate and phishing websites.
- Minimal misclassification.

Explainable AI (XAI) Techniques

1. Feature Importance (GBM Inherent)

- Highlights features directly used by the model during training.
- **Top 5 Features**:
  1. SSLfinal_State : 0.695
  2. URL_of_Anchor : 0.144
  3. Prefix_Suffix : 0.04
  4. Total_Link_Flags : 0.0397
  5. Web_traffic :0.0187

Interpretation:These features, according to the Gradient Boosting model, are the most relevant with regard to the prediction of phishing likelihood.

---

2. SHAP (SHapley Additive exPlanations)

- **Type**: Model-specific (TreeExplainer)
- Offers global and local interpretability.
- **SHAP Summary Plot**:
  - Visualizes the distribution and impact of each feature on predictions.
  - Feature values are color-coded to reveal direction of influence.

---

3. LIME (Local Interpretable Model-Agnostic Explanations)

- **Type**: Local, model-agnostic
- Applied to a random test instance.
- Produces interpretable linear approximation around a single prediction.
- **Output**:
  - Presents ten most important features that contributed to the prediction.
  - Explanations justify the decision by showing the weights and directions of features.

---

4. PDP & ICE (Partial Dependence & Individual Conditional Expectation)

- **Feature**: SSLfinal_State
- **Insight**:
  - Indicates the way in which predictions generated by the model vary with the final state of the SSL.
  - ICE curves show different effects for each individual instance while PDP shows the average effect.
  - Higher SSLfinal_State values are associated with an increase in phishing probability.

---

5. PFI (Permutation Feature Importance)

- **Type**: Model-agnostic, global

- Evaluates how random shuffling of each feature impacts model performance.
- **Top PFI Features**:
  - Consistent with GBM feature importance and SHAP but some difference in the last two features.
  - Further validates the importance of SSLfinal_State and URL_of_Anchor.

Conclusion

The Gradient Boosting model does a wonderful job in recognizing phishing websites. The interpretability techniques that one could apply in the context of this model include SHAP, LIME, PDP/ICE, and PFI. These methods not only validate the model's decisions but also justify to the user the reasons that cause his or her outputs. This is a very important step toward the trust and transparency of AI-based cyber security systems.