Explainability Report: Logistic Regression for Phishing Website Detection

---

Model Overview

- **Model**: Logistic Regression
- **Target**: Binary classification (1 = Phishing, -1 = Legitimate)
- **Data**: UCI Phishing Website Dataset (with engineered features)

---

1. Logistic Regression Assumptions Check

Assumption	Status	Notes
Binary outcome	✅ Satisfied	Target values: 1, -1
Independence of observations	✅ Assumed (given no time series or grouping)	
Linear relationship (logit link)	✅ Approximate via feature relationships	Visualized: Actual vs Predicted
Sample size adequacy	✅ Satisfied	~10+ samples per feature/class
No multicollinearity	⚠️ Partial Violation	Several VIF > 10

VIF (Variance Inflation Factor) revealed potential multicollinearity. While no *definite* multicollinearity (VIF > 100) was found, several variables exceeded VIF > 10, indicating possible redundancy and affecting interpretability of coefficients.

---

3. Feature Interpretability

Feature Importance via Coefficients

Feature	Coefficient (impact)
`statistical_report`	**+ve** — strong indicator of phishing
`https_token`	+ve
`web_traffic`	-ve
`having_at_symbol`	+ve
`request_url`	+ve

Positive coefficients imply higher likelihood of phishing when feature is present/active.

---

4. Model Explainability

SHAP (SHapley Additive Explanations)

- **Tool Used**: `shap.LinearExplainer` (optimized for linear models)

- **Sample Size**: 10 test samples with background of 100 training instances

SHAP Summary Plot Insights:

- Top influencers:
  - `https_token`
  - `statistical_report`
  - `having_at_symbol`
  - `request_url`
  - `obfuscation_score`
- SHAP values clearly show whether each feature **increased** or **decreased** phishing probability.

LIME (Local Interpretable Model-agnostic Explanations)

- Explained one random instance using `lime_tabular`.
- Output: Top 10 features influencing a single prediction.
- Displayed weights for both phishing and legitimate classes.