Extended Explainability Report: Neural Network Model – MLP Classifier

Model Summary

Model	Description
MLP Classifier (MLP)	Multi-layer Perceptron — fully connected neural network capable of capturing complex, nonlinear feature interactions. Optimized with backpropagation.

Model Performance

Metric	Value
Accuracy	95%

Comments:

- Performs well with nonlinear decision boundaries.
- Slightly longer training time due to backpropagation over 1000 iterations.
- May be sensitive to feature scaling and hyperparameter tuning.

Explainability with SHAP

MLP SHAP Summary

- **SHAP KernelExplainer** used on `.predict()` approximation.
- **Top Influencers** (Global Feature Importance):
  - `URL_of_Anchor`
  - `SFH`
  - `Prefix_Suffix`
  - `Request_URL`
  - `web_traffic`

Insights:

- SHAP values show how individual features contribute to neural network outputs.
- Nonlinear interactions are partially interpretable using SHAP approximations.
- Visual summary plots highlight key drivers for both phishing and legitimate predictions.

Explainability with LIME

MLP Classifier (LIME)

- Local explanation generated for a random test instance.
- **Top Local Influencers**:
  - `Request_URL`
  - `having_Sub_Domain`

  - `URL_of_Anchor`
  - `web_traffic`

Insights:

- LIME complements SHAP by offering case-by-case reasoning.
- Confirms that certain features (e.g., `Request_URL`) consistently push predictions toward phishing.

Permutation Feature Importance (PFI)

Rank	Feature	Importance
1	URL_of_Anchor	High
2	Prefix_Suffix	High
3	SFH	Moderate
4	web_traffic	Moderate
5	Request_URL	Moderate

Insights:

- Permutation tests confirm model dependence on high-impact URL structure indicators.
- Shuffling these features significantly reduces model accuracy.

Leave-One-Feature-Out (LOFO) Importance

Rank	Feature	Accuracy Drop
1	Prefix_Suffix	Highest
2	URL_of_Anchor	High
3	SFH	High
4	web_traffic	Moderate
5	having_Sub_Domain	Moderate

Insights:

- Removing individual features reveals their **critical importance** to model performance.
- LOFO complements SHAP by highlighting **essential dependencies**.