Explainable AI for Financial Machine Learning Models

Introduction

In financial domains, machine learning (ML) models are often used for critical tasks such as credit approval, fraud detection, and risk assessment. These models, especially complex ones like gradient boosting or deep neural networks, are often perceived as *black boxes* because their internal decision-making logic is not directly interpretable by humans.

Problem of Interpretability in Finance

Financial institutions operate under strict regulatory standards and require high levels of transparency to build trust with clients and regulators. A lack of interpretability can lead to:

- Reduced trust in model predictions
- Legal and compliance issues
- Difficulty in debugging or improving models

Why SHAP and LIME Help

SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are two prominent model-agnostic techniques designed to explain individual predictions.

- **SHAP**: Based on cooperative game theory (Shapley values), it assigns an importance value to each feature by considering all possible combinations of feature contributions.
- **LIME**: Perturbs the input data locally and trains an interpretable surrogate (usually linear) model to approximate the complex model's behavior near a specific prediction.

Algorithm Outline

SHAP Algorithm Steps

1. Select an instance (example) for which an explanation is needed.
2. Compute the model output with and without each feature, across all possible subsets (coalitions).
3. Average the marginal contributions of each feature across all subsets to calculate the Shapley value.
4. Aggregate these values into a per-feature explanation vector.
5. Visualize or report these values to understand feature impact.

LIME Algorithm Steps

1. Choose an instance to explain.
2. Generate perturbed samples around this instance.
3. Compute predictions for these perturbed samples using the black-box model.
4. Fit a simple interpretable model (e.g., linear regression) to these samples.
5. Extract coefficients from the interpretable model as feature attributions.

Conclusion

Using SHAP or LIME enables financial practitioners to explain predictions at an individual level, improve model transparency, and satisfy compliance requirements. These tools empower model developers and auditors to understand the *why* behind predictions, not just the *what*.