Digital Finance 3: Technology in Finance

Lesson 35: Explainability and Bias

FHGR

December 13, 2025

Summary of key concepts presented above.

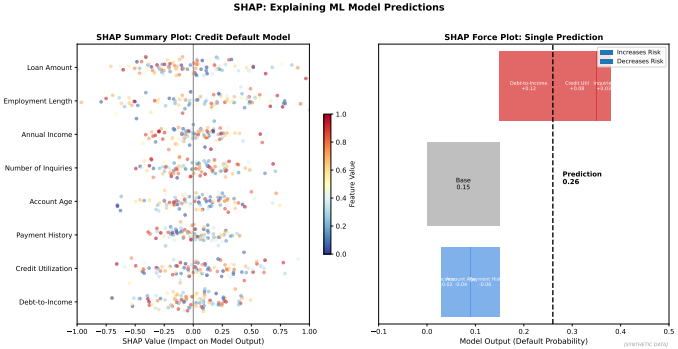
Learning Objectives

By the end of this lesson, you will be able to:

- Explain the interpretability-accuracy trade-off
- Apply SHAP and LIME for model explanations
- Understand feature attribution methods
- Detect and mitigate algorithmic bias
- Evaluate fairness metrics in financial ML
- Navigate regulatory requirements (GDPR Article 22)

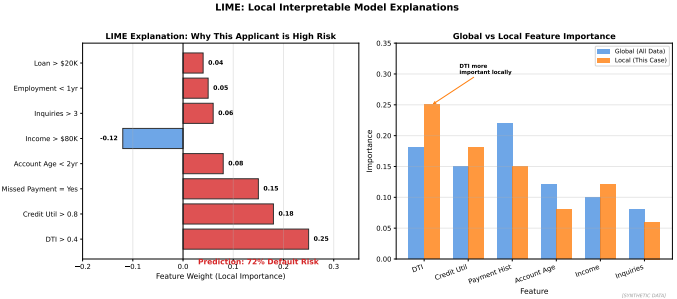
Summary of key concepts presented above.

SHAP Values for Feature Importance



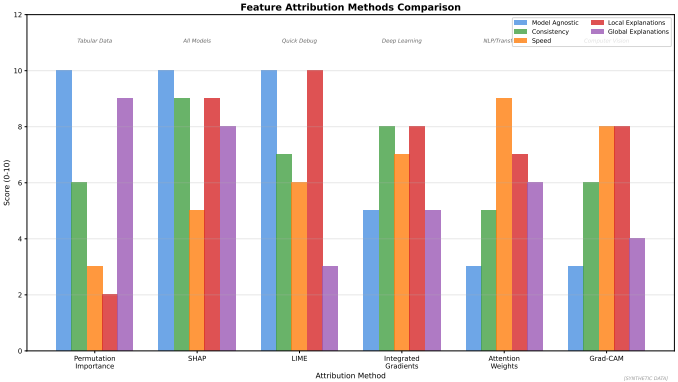
SHAP values decompose predictions into individual feature contributions based on game theory.

LIME: Local Interpretable Model-Agnostic Explanations



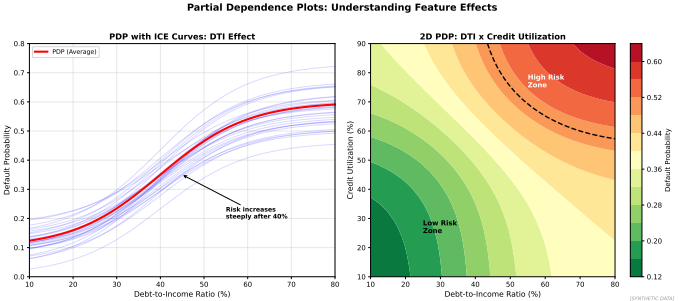
LIME approximates black-box models locally with interpretable linear models for individual predictions.

Feature Attribution Methods Comparison



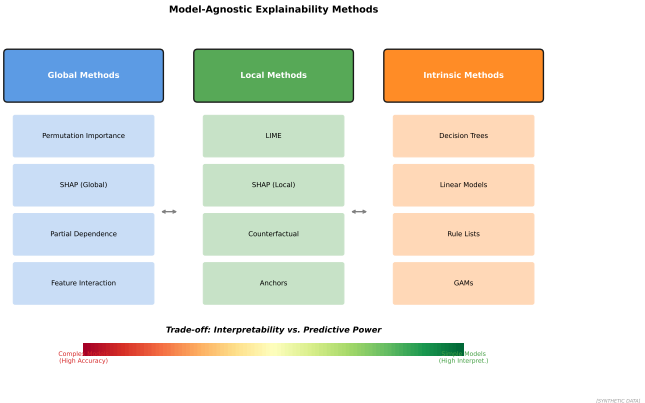
Different attribution methods provide complementary insights into model behavior and feature importance.

Partial Dependence and ICE Plots



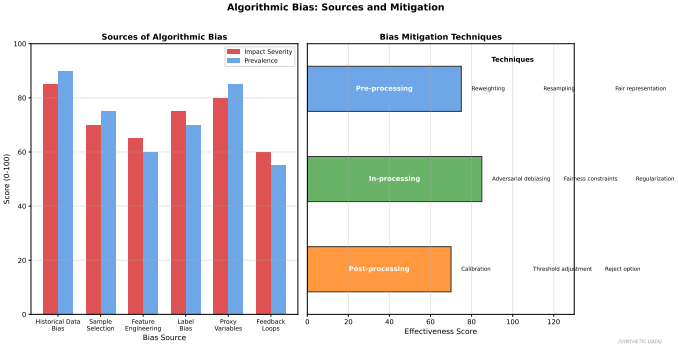
PDP shows average marginal effects; ICE plots reveal heterogeneous effects across instances.

Model-Agnostic Explainability Methods



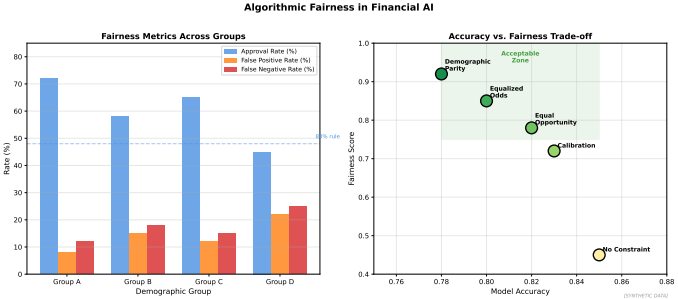
Model-agnostic methods work with any ML model, enabling consistent explanations across model types.

Algorithmic Bias Sources



Bias can arise from training data, feature selection, model design, or deployment decisions.

Fairness Metrics Comparison



Multiple fairness definitions exist; choosing the right metric depends on context and stakeholder values.

Summary

Key Takeaways:

- Explainability required by regulations (GDPR Article 22)
- SHAP and LIME most popular explanation methods
- Trade-off: accuracy vs. interpretability
- Bias detection critical for fair lending and hiring
- Multiple fairness metrics (no one-size-fits-all)
- Explainability tools maturing rapidly

Next Lesson: AI Regulation and Future

Summary of key concepts presented above.