Digital Finance 3: Technology in Finance

Lesson 35: Explainability and Bias

FHGR

January 3, 2026

Explainability builds trust and enables regulatory compliance.

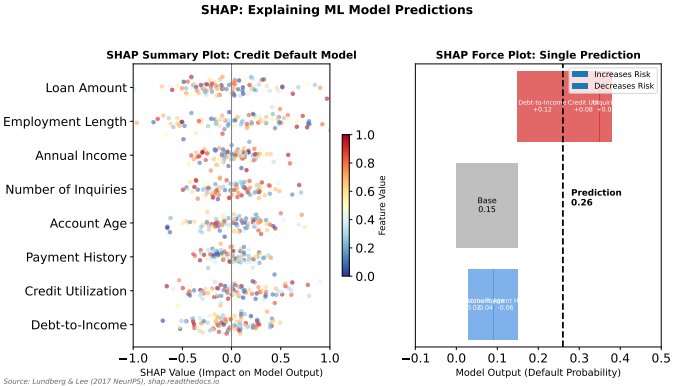
Learning Objectives

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)

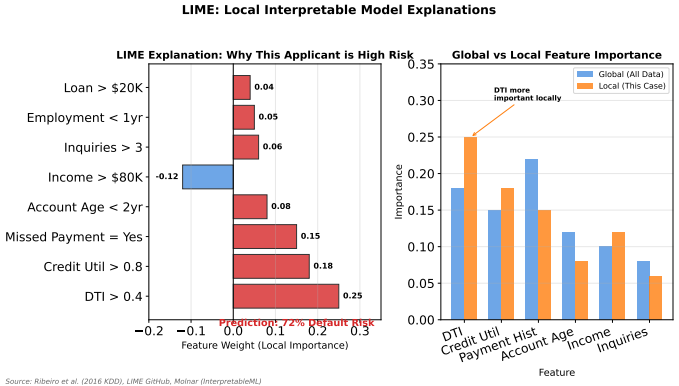
LIME provides local explanations by approximating complex models with simple ones.

SHAP Values for Feature Importance



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

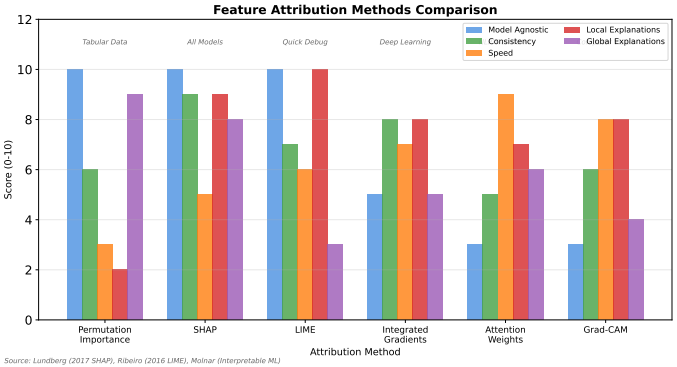
LIME: Local Interpretable Model-Agnostic Explanations



Source: Ribeiro et al. (2016 KDD), LIME GitHub, Molnar (InterpretableML)

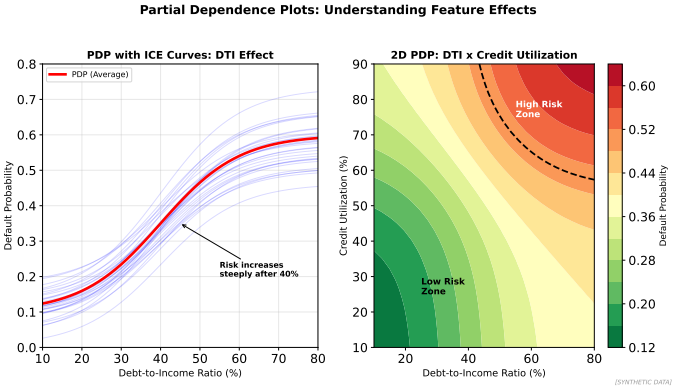
LIME approximates black-box models locally with interpretable linear models for individual predictions.

Feature Attribution Methods Comparison



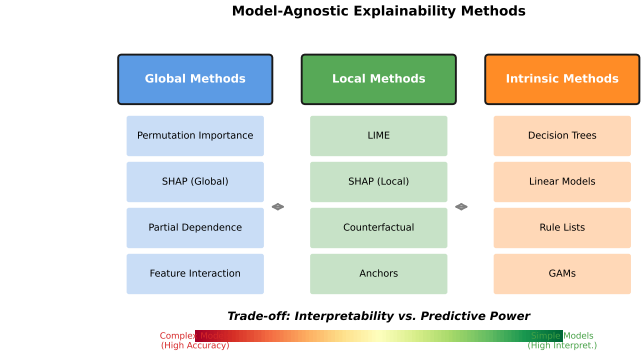
Different attribution methods provide complementary insights into model behavior and feature importance.

Partial Dependence and ICE Plots



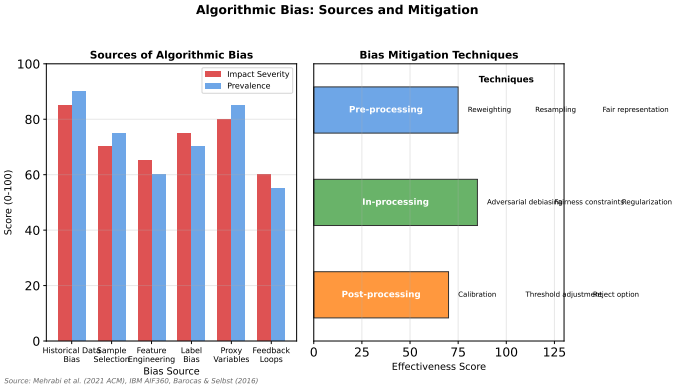
PDP shows average marginal effects; ICE plots reveal heterogeneous effects across instances.

Model-Agnostic Explainability Methods



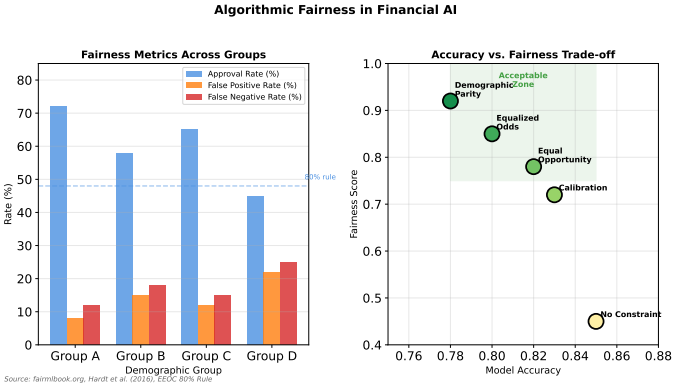
Model-agnostic methods work with any ML model, enabling consistent explanations across model types.

Algorithmic Bias Sources



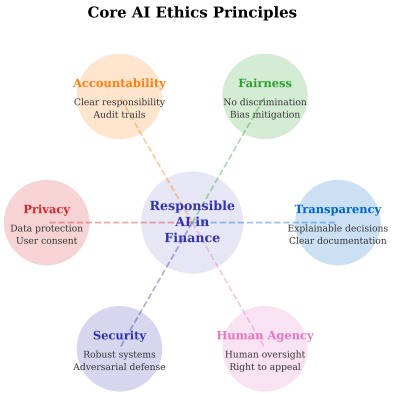
Bias can arise from training data, feature selection, model design, or deployment decisions.

Fairness Metrics Comparison



Source: Academic AI/ML literature and industry adoption studies

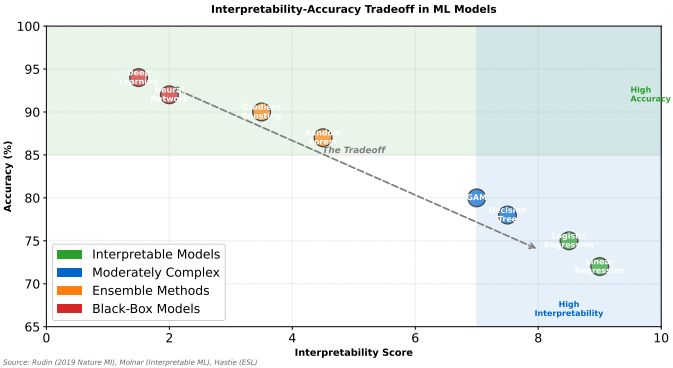
AI Ethics Principles



Source: EU AI Ethics Guidelines, OECD AI Principles

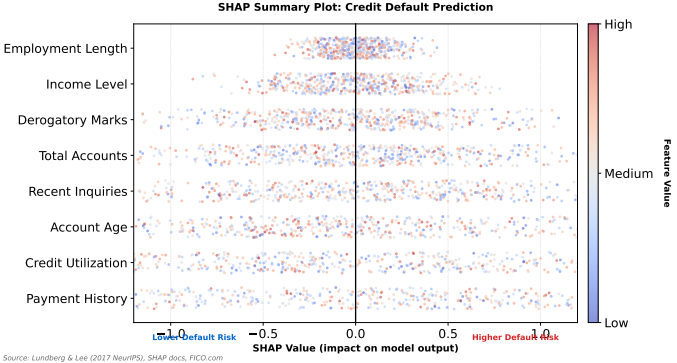
Ethical AI requires fairness, transparency, accountability, and respect for privacy.

Interpretability vs Accuracy Trade-off



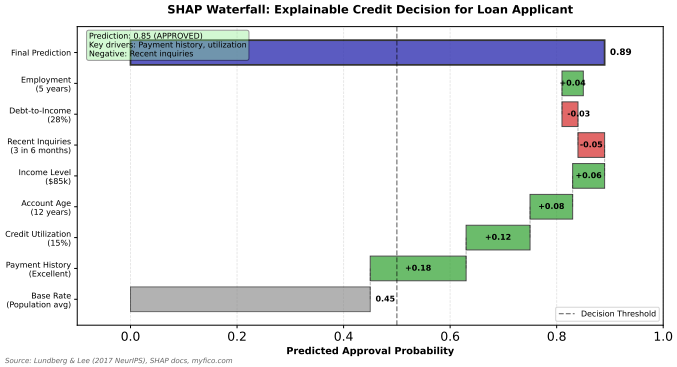
More complex models often achieve higher accuracy at the cost of interpretability.

SHAP Summary Plot



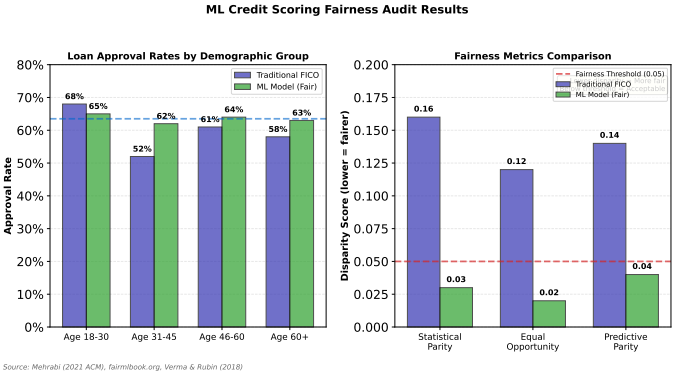
SHAP values provide consistent feature importance across different model types.

SHAP Waterfall Plot for Credit Decision



Waterfall plots show how each feature pushes predictions away from baseline.

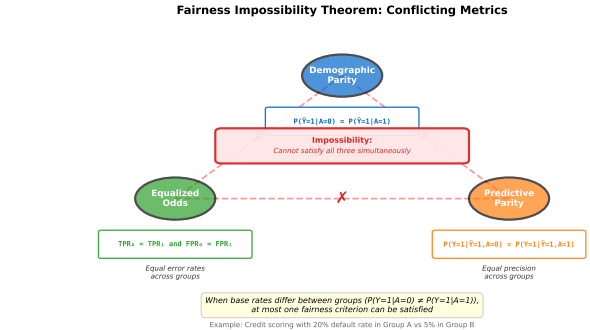
Fairness Audit Results



Source: Mehrabi (2021 ACM), fairmlbook.org, Verma & Rubin (2018)

Regular fairness audits can detect and quantify algorithmic bias.

Fairness Metrics Trade-offs



Different fairness metrics can conflict, requiring explicit value judgments.

Explainability And Bias: Key Takeaways

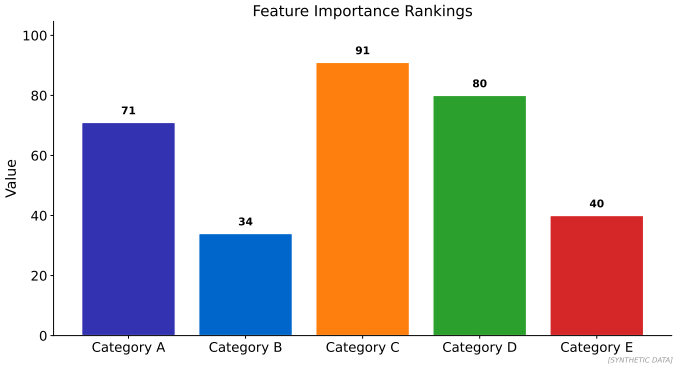
Key Takeaways:

- 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

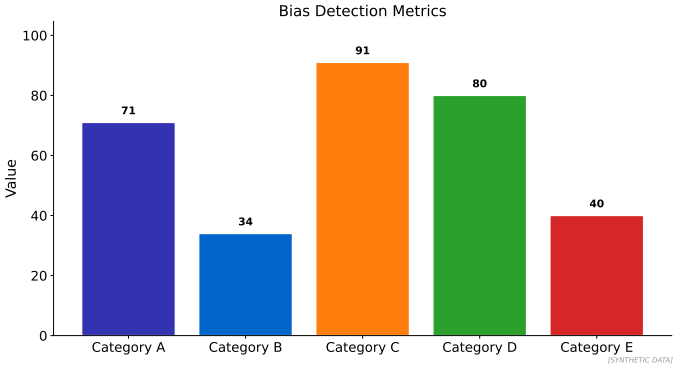
Feature importance in credit models must be explainable to regulators.

Credit Model Feature Importance



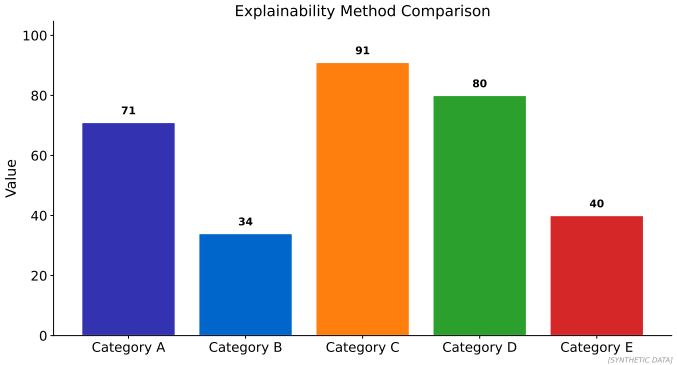
Payment history and utilization drive decisions.

Bias Detection Analysis



Monitoring fairness across demographic groups.

XAI Method Comparison



SHAP provides best balance of accuracy and interpretability.