XAI for greener portfolio decisions

Prof. Dr. Audrius Kabašinskas
Lecturer PhD student Jurgita Černevičienė

MSCA Digital Finance – Training week on Green Digital Finance

KTU, 30 June – 04 July 2025

Introduction to XAI and Sustainable Finance

- Explainable Artificial Intelligence (XAI) and its importance in finance
- The concept of greener portfolio decisions and sustainable finance (ESG investing, climate finance)
- The challenge of "black-box" AI models and the need for explainability in sustainable investment decisions

Outline

What is Explainable AI (XAI)?

- AI models often act as **"black boxes"** - decisions are made without clear reasoning visible to users.
- XAI provides transparency by explaining **how** AI models arrive at their decisions.
- Allows humans to **understand**, trust, and manage AI-driven outcomes.

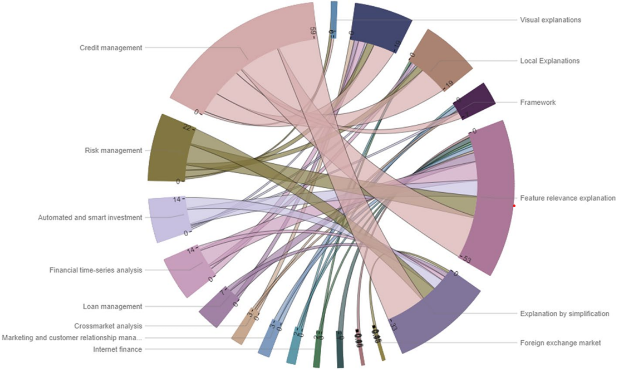
Why XAI Matters in Finance

- **Regulatory Compliance**: Laws (e.g., EU AI Act, GDPR, US ECOA) require transparency in automated financial decisions.
- **Trust & Accountability**: Stakeholders need to understand the AI decisions that affect loans, investments, and fraud detection.
- **Bias Detection**: XAI helps identify and mitigate unfair biases in credit scoring and risk assessment.

Key Use Cases of XAI in Finance

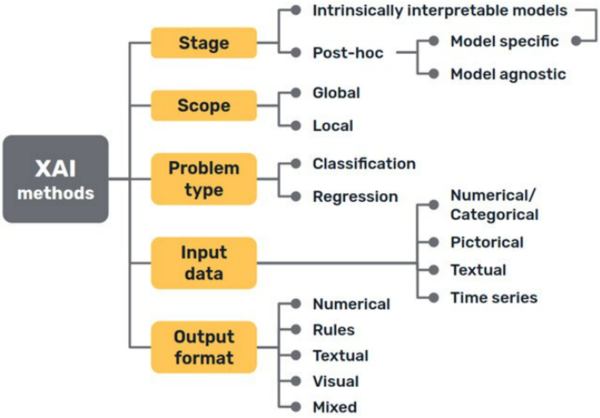
- **Credit Scoring**: Explaining why a loan was approved or denied
- **Fraud Detection**: Identifying suspicious transaction patterns with reasoning
- **Investment Recommendations**: Transparent rationale for portfolio suggestions
- **Insurance**: Clear explanations for premium calculations
- **Risk Assessment**: Interpretable models for market and credit risk
- **Algorithmic Trading**: Understanding automated trading decisions
- **Customer Service**: Explaining AI-driven responses and recommendations

Key Use Cases of XAI in Finance



Further reading: Černevičienė and Kabašinskas (2024) - Explainable artificial intelligence (XAI) in finance: a systematic literature review.

How XAI Works — Methods

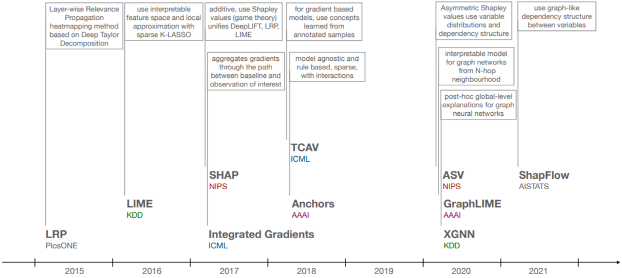


How XAI Works — The Five Dimensions of XAI Methods

- **Stage of Application**
  - Intrinsically Interpretable Models
  - Post-hoc Explanations
  - Model-Specific Techniques
  - Model-Agnostic Techniques
- **Scope of Explanation** assess absolute performance without reference.
  - Global Explanations
  - Local Explanations
- **Problem Type**
  - Classification Problems
  - Regression Problems
- **Input Data Types**
- **Output Format**

XAI Techniques

Further reading: Holzinger et al. (2021) - Explainable AI Methods - A Brief Overview.



XAI Techniques

SHAP (Shapley Additive Explanations)

To calculate the Shapley value ϕ_i of a specific feature i, we find the average prediction difference between the model with feature i and the model without it (Lundberg and Lee (2017)). This is done over all feature subsets $S \subseteq M \setminus \{i\}$, considering all feature orderings. The expression that defines SHAP values is as follows (Mitchell (2022)):

$$\phi_i = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(|M| - |S| - 1)!}{|M|!} \left[f_{S \cup \{i\}}(x) - f_S(x) \right]$$

M – the set of all features $f_S(x)$ – the model output restricted to the feature subset S x – a feature vector

GitHub Repo: https://github.com/slundberg/shap

XAI Techniques

LIME (Local Interpretable Model-Agnostic Explanations)

This method approximates the black-box model locally around an individual instance rather than explaining it globally Ribeiro et al. (2016)). The explanation $\xi(x)$ is obtained by solving the following optimization problem:

$$\xi(x) = \arg\min_{g \in \mathcal{G}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$\mathcal{G}$ – the set of interpretable models, f – the black-box model, g – an interpretable model, π_x – a proximity measure around instance x, $\mathcal{L}(f, g, \pi_x)$ – loss function measuring how well g approximates f near x, $\Omega(g)$ – complexity penalty for model g
GitHub Repo: https://github.com/marcotcr

XAI Techniques: Anchors (1/2)

Anchors explain individual predictions of black-box models using *IF-THEN decision rules* that "anchor" a prediction to a specific class (Ribeiro (2018)). **Core Idea:**

- Anchors define **regions in feature space** where the model prediction remains unchanged.
- Within these regions, predictions stay **fixed** even if non-anchor features vary.
- Designed to produce **local explanations** that are simple and human-readable.

Model-Agnostic: Can be applied to *any* predictive model without access to internal structure.

XAI Techniques: Anchors (2/2)

Anchor Properties:

- **Precision**: Proportion of data points satisfying the anchor that share the target class.
- **Coverage**: Proportion of data points to which the anchor rule applies.
- A good anchor should have both **high precision** and **high coverage**.

How Anchors Are Found:

- Uses **reinforcement learning (RL)** and a modified **beam search** algorithm.
- Generates and extends **candidate rules** iteratively.
- Treats anchor selection as a **multi-armed bandit** problem due to repeated model queries.

GitHub Repo:https://github.com/marcotcr/anchor

XAI Techniques:Counterfactual Explanations (1/2)

A **counterfactual explanation** answers the 'what if' question: if X had not occurred, Y would not have occurred. Describes individual predictions by suggesting minimal changes in characteristics that produce a predefined result:

- Define a target outcome (e.g., loan approval instead of rejection).
- Find the smallest change to the input x—denoted x'—that causes the model $\hat{f}$ to yield the desired output.

Examples:

- Loan scenario: If Peter's income were €10,000 higher, the model would accept his application.
- Rent prediction: If the apartment were 15m² larger or if pets were allowed and windows insulated, the predicted rent reaches €1000.

XAI Techniques:Counterfactual Explanations(2/2)

How Counterfactuals Are Generated:

- **Optimization (OPT)**: Counterfactual explainers based on optimization strategies define a loss function that accounts for desired properties and adopts existing optimization algorithms to minimize it.
- **Heuristic Search Strategy (HSS)**: aim to find counterfactuals through local and heuristic choices that at each iteration minimize a certain cost function.
- **Instance-Based (IB)**: retrieve counterfactuals by selecting the most similar examples from a dataset.
- **Decision Tree-Based (DT)**: approximate the behavior of the black-box with a decision tree and then exploit the tree structure to identify counterfactual explanations.

GitHub Repo:
https://christophm.github.io/interpretable-ml-book/counterfactual.html

XAI Techniques: Partial Dependence Plot (PDP)

Partial Dependence Plot (PDP) visualizes the marginal effect of one or more features on a model's predicted outcome.

Key Properties:

- Model-agnostic: applicable to any machine learning model.
- Shows whether a feature's relationship with the target is linear, monotonic, or complex.
- Useful for validating model behavior against domain knowledge.

How PDP Works:

1. Select a feature and define a grid of its values.
2. Replace the feature with each grid value across all instances.
3. Predict outcomes and compute the average prediction.
4. Plot the grid values against these average predictions.

Limitation: PDP does not capture feature interactions and assumes independence between features.

Benefits of XAI in Finance

- Enhances the quality and transparency of decision making.
- Builds stronger relationships with customers and regulators.
- Supports auditability and compliance with evolving regulations.
- Enables proactive risk management and bias mitigation.

Outline

The Challenge of Greener Portfolio Decisions

- **Sustainable investing** requires balancing financial returns with environmental, social and governance (ESG) goals.
- AI models help to analyze complex ESG data but often act as black boxes that are **difficult to interpret**.
- Lack of transparency hinders trust, regulatory compliance, and stakeholder acceptance.

XAI Enhances Transparency and Trust

- XAI explains how AI models make decisions, revealing key ESG factors that influence portfolio choices
- Bias Detection - Identifies unfair patterns and enables proactive correction
- Builds investor and regulator confidence by clarifying rationale behind sustainable investments
- Promotes accountability and ethical finance aligned with green objectives

Improving Risk Management and Bias Mitigation

- XAI helps detect and mitigate biases in ESG data and AI models, ensuring fair treatment of all sectors and communities.
- Enables dynamic assessment of climate and biodiversity risks with interpretable insights.
- Supports proactive portfolio adjustments in response to evolving sustainability trends
- **Further reading:** van der Heever W et al. (2024) -Understanding Public Opinion towards ESG and Green Finance with the Use of Explainable Artificial Intelligence

Facilitating Regulatory Compliance and Reporting

- Explainability **supports adherence** to sustainability regulations (e.g., SFDR (Sustainable Finance Disclosure Regulation), EU Taxonomy, TCFD (Task Force on Climate-related Financial Disclosures)).
- **Simplifies auditing and reporting** by providing clear evidence of ESG risk assessments and investment decisions.
- Ensures portfolios meet evolving disclosure and compliance **requirements**.
- Enables **Real-Time Compliance Monitoring**: Explainable systems can flag non-compliance risks early by revealing when portfolios drift from regulatory thresholds or ESG objectives.
- **Mitigates Legal and Reputational Risk**. By making investment decisions explainable, organizations can defend against green-washing claims and demonstrate good-faith efforts to comply with sustainable investing mandates.

Driving Better Portfolio Performance through Explainability

- **Enhances Investment Decisions:** Clarifies how ESG and financial factors drive performance, enabling more informed and confident portfolio choices.
- **Improves Risk-Adjusted Return:** Helps balance sustainability goals with financial performance by identifying the true drivers of risk and return.
- **Enables Proactive Risk Management:** Makes ESG and climate-related risks more visible and actionable, supporting early intervention.
- **Accelerates Strategy Optimization:** Facilitates quicker iteration and refinement of investment models by revealing what's working and why.
- **Builds Trust and Client Engagement:**Transparent explanations improve stakeholder confidence and support stronger, long-term client relationships

Case Studies of XAI

- Çankal, A., Ever, D. (2025). The Effects of Renewable Energy Consumption on Financial Performance: An Explainable Artificial Intelligence (XAI)-Based Research on the BIST Sustainability Index. International Journal of Energy Economics and Policy, 15(4), 204-213. XAI can quantify renewable energy's financial impacts, revealing positive correlation between sustainability practices and financial performance despite modest immediate returns.
- Saxena, A., Santhanavijayan, A., Shakya, H. K., Kumar, G., Balusamy, B., Benedetto, F. (2024). Nested Sentiment Analysis for ESG Impact: Leveraging FinBERT to Predict Market Dynamics Based on Eco-Friendly and Non-Eco-Friendly Product Perceptions with Explainable AI. Mathematics, 12(21), 3332. This study creates a transparent, high-accuracy framework for incorporating market sentiment about environmental products into ESG scoring, bridging the gap between public perception of sustainability and formal ESG metrics through explainable AI methods.
- Radzkova, K. (2024). Transparency and Efficiency in Credit Risk Assessment of Alternative Financing: A Green AI Approach (Bachelor's thesis, University of Twente). Shows how explainability and environmental responsibility can be integrated into practical financial applications

References:

- Černevičienė, J., Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artificial Intelligence Review, 57(8), 216.
- Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W. (2020, July). Explainable AI methods-a brief overview. In International workshop on extending explainable AI beyond deep models and classifiers (pp. 13-38). Cham: Springer International Publishing.
- Ribeiro, M. T., Singh, S., Guestrin, C. (2018, April). Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).
- Stiefenhofer, P., Deniz, C., Chen, Y., Qian, J., Almehthel, H. M. (2024). The Future of Sustainable Finance: AI-Driven Sustainable Pairs Trading in Market-Neutral Investing. In Artificial Intelligence, Finance, and Sustainability: Economic, Ecological, and Ethical Implications (pp. 111-142). Cham: Springer Nature Switzerland.