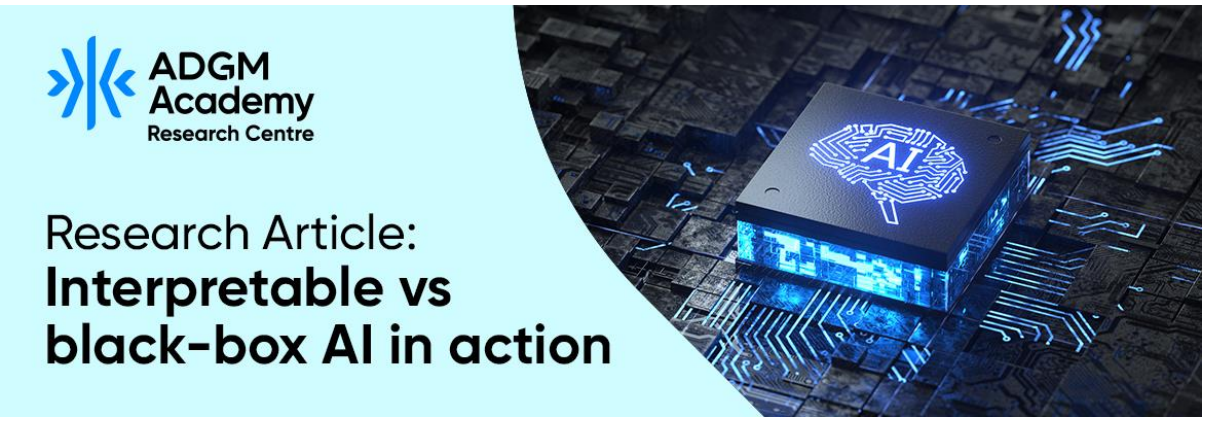
ADGM Academy Research Centre

Article SERIES



INTERPRETABLE VS BLACK-BOX AI IN ACTION

Artificial Intelligence (AI) has made significant strides in recent years, but its adoption in managerial and policy contexts faces hurdles due to the "black-box" nature of many models like neural networks. Interpretability is a critical requirement in industries like finance, healthcare, and regulatory environments, where decision-makers need to understand and justify model outputs.

Jin-Chuan Duan is a member of the ADGM Academy Research Centre Advisory Board, and Chairman of ADBIZA and Criat. He is also professor emeritus of the National University of Singapore and currently serves as an adjunct chair professor in the College of Global Banking and Finance, National Chengchi University.

His recent paper, *Interpretable vs Black-box AI in Action*, highlights a machine learning approach that enhances traditional interpretable models with modern optimization techniques, offering comparable or superior performance to black-box AI on tabular data.

The paper presents two case studies:

- Hedonic regression for real estate pricing.
- Vector autoregression (VAR) for macroeconomic forecasting.

Both examples demonstrate how interpretable AI can meet performance demands without sacrificing clarity.

WHY INTERPRETABILITY MATTERS

Black-box models, such as neural networks, are powerful but lack transparency. This limitation is particularly problematic in industries where accountability, regulations, and tradition demand decision-makers understand and trust model outputs. Interpretability provides clarity and builds confidence, enabling:

- **Extrapolation**: Confidence in predictions beyond training data.
- **Legal and regulatory compliance**: Models must withstand scrutiny in highly regulated sectors.

- User acceptance: Stakeholders prefer models grounded in theory or common sense.

Interpretability Defined

The paper distinguishes between:

- Explainability: Explaining how a model reaches its predictions (e.g., random forests).
- Interpretability: Models that users can intuitively understand and trust (e.g., decision trees, hedonic regressions).

METHODOLOGY: ENHANCED INTERPRETABLE MODELS

The proposed approach improves traditional models by:

- Using combinatorial optimization to identify the most relevant features from high-dimensional datasets, irrespective of the number of data instances.
- Ensuring models remain parsimonious (simple and efficient) while maintaining high performance.

Key Techniques

- Sequential Monte Carlo (SMC) Optimization: Efficiently searches for optimal feature combinations across large-dimensional datasets.
- Stable Optimized Decision Trees (SODT): Groups categorical variables into meaningful clusters for better interpretability.
- Stable Combinatorially Optimized Feature Selection (SCOFS): Selects a minimal but effective subset of variables.

CASE STUDIES

1. Real Estate Pricing with Hedonic Regression

Using the Ames, Iowa housing dataset, the authors developed an interpretable hedonic pricing model that predicts house prices based on features like size, location, and quality.

Key Results

Two-stage modelling process:

- **Stage 1**: Grouped similar categorical features such as neighbourhoods into 10 composite clusters, achieving an R^2 of 72.96%.
- **Stage 2**: Added interaction terms for greater precision, achieving R^2 values of 93.89% (training data) and 92.16% (testing data).

Comparison with Black-box Models:

- Outperformed random forests and neural networks on test data.
- Maintained interpretability with insights into feature importance (e.g., "Gross Living Area" contributes $47,840 per 1,000 square feet).

Practical Implications: The interpretable model allows real estate professionals to confidently use it for extrapolation, even for properties outside the training data.

2. Macroeconomic Forecasting with VAR

The second case study developed a parsimonious VAR model to analyze the relationships between seven macroeconomic variables, including consumption growth, inflation, and the Fed Funds rate.

Key Results

- Reduced model complexity from 812 potential coefficients to just 21 using SCOFS.
- Achieved stable and interpretable results, highlighting intuitive relationships (e.g., Fed Funds rate responding to investment growth during specific economic periods).

Practical Implications: Economists and policymakers can use this model to forecast and analyze economic trends with greater clarity and trust.

COMPARATIVE PERFORMANCE

The interpretable models demonstrated competitive performance compared to black-box approaches:

- Real Estate Pricing Model:
  - R^2: 92.16% (testing data) vs. 91.29% (neural networks with two hidden layers).
  - Median pricing error: 6.85%.
- Macroeconomic VAR Model:
  - Achieved a parsimonious structure that is easier to interpret and validate.

BROADER IMPLICATIONS FOR INDUSTRY

The research showcases a pathway to bridge the gap between cutting-edge AI and the practical needs of industry:

- **Compliance and Accountability**: Models align with regulatory requirements and stakeholder expectations.
- **Adoption and Trust**: Interpretability fosters greater acceptance among non-technical users.
- **Scalability**: The optimization approach can enhance a wide range of conventional models.

CONCLUSION

The increasing reliance on Artificial Intelligence in decision-making processes across various industries highlights the importance of model interpretability, especially in managerial and policy-driven contexts. While black-box models like neural networks and random forests have demonstrated exceptional predictive power, their opaque nature limits their utility in environments that demand accountability, compliance, and user trust. Interpretability is not merely a desirable feature but a necessity in sectors such as finance, healthcare, and real estate, where decisions must be transparent, defensible, and aligned with established regulatory frameworks.

This paper illustrates that traditional, interpretable models can be significantly enhanced through modern machine-learning techniques, such as sequential Monte Carlo combinatorial optimization. By leveraging these advancements, conventional models can achieve high performance while retaining their inherent transparency and simplicity. The hedonic regression model for real estate pricing and the parsimonious vector autoregression (VAR) for macroeconomic forecasting exemplify this approach, demonstrating how interpretable AI can not only meet but sometimes exceed the performance of black-box alternatives.

The findings have broad implications for industry practitioners. By adopting interpretable AI, organizations can better align their analytical capabilities with stakeholder expectations,

enabling confident decision-making and compliance with stringent regulations. Furthermore, interpretable models allow for meaningful extrapolation and intuitive insights, empowering users to navigate complex scenarios with ease. The research also underscores the role of machine-learning innovations in bridging the gap between theoretical rigor and practical usability.

Looking ahead, the challenge lies in scaling these techniques to address increasingly large and complex datasets without compromising interpretability. As computational power continues to grow and optimization algorithms become more sophisticated, the potential for interpretable AI to become the default choice for managerial applications is immense. This paper advocates for a paradigm shift: instead of forcing interpretability onto black-box models, the focus should be on refining and enhancing traditional models to meet modern data demands. By doing so, organizations can harness the power of AI in ways that are not only effective but also comprehensible, fostering trust and long-term adoption across industries.

ADGM Academy
Research Centre

Follow / Contact Us:

www.adgmacademy.com

research@adgm.com

LinkedIn