Model interpretation: It’s not what it looks like

Abstract

Model interpretation techniques have become crucial for the adoption of machine learning algorithms in the credit industry. We rely on interpretation techniques to justify model behavior, ensure fairness, and derive business insights. However, these interpretation methods do not always provide the full picture, especially for complex models.

In this talk, we will explore the practical challenges of model interpretation and discuss why conventional techniques might be misleading or insufficient. We will assume the audience has prior knowledge of ML interpretation methods and will begin with a brief recap of common techniques such as feature importance (Split, Permutation FI, SHAP, dropout) and feature interpretation (PDP, LIME, SHAP). Rather than diving into the technical mechanics, we will only describe their high-level concepts for the context of following discussions.

We will then shift our discussion to a broader perspective on interpretation, drawing an analogy to understanding human decision-making. We will highlight how the effectiveness of interpretation depends on:

1. Different purposes
a. Understanding model mechanics
b. Justifying causal relationships
c. Comparative analysis
d. Extracting actionable insights

2. Different contexts
a. Individual vs. population-level interpretation
b. Choice of reference points
c. Features within the model vs. left out

3. Different levels of precision required
a. Order of magnitude estimation
b. Directional interpretation
c. Edge cases

Finally, we will address technical challenges and practical caveats in popular interpretation techniques:

1. The impact of correlated feature sets on interpretability
2. Challenges in interpreting stacked models

By the end of this talk, attendees will gain a deeper understanding of the complexities behind model interpretation and leave with a more nuanced perspective on when and how to trust interpretation techniques in real-world applications.

Authors & Affiliations

Dr Jiahang Zhong[1]
[1]Monzo Bank, London, United Kingdom