# Enhancing Model Explainability

Gael Decoudu

23/05/2024

# Interpretability or Explainability?

- **Interpretability** refers to **the ability to understand how a model works** internally and how it arrives at its predictions or decisions. An interpretable model has transparent internal mechanics that can be comprehended and reasoned about by humans. This is typically achieved by using inherently **interpretable models such as linear regression, decision trees, or rule-based systems.**
- **Explainability** refers to the ability to **provide explanations for a model's predictions or decisions after the fact**, even if the internal mechanics of the model are not fully interpretable. Explanations can be generated using various techniques, even **for complex "black-box" models like neural networks or ensemble models.**

Guidotti et al. (2018) in their survey paper on explainable AI state: "Interpretability is the ability to explain or to present in understandable terms to a human." "Explainability refers to the details and reasons a machine can provide to make its functioning understandable to humans."

# The importance of model explainability in financial services

- **Regulatory requirements for transparency**
  - **GDPR:** Article 22 provides individuals with the "right to explanation" for decisions made by automated systems that significantly affect them. This highlights the need for explainable AI systems.
  - **EU AI Act:** High-risk AI systems must be designed with techniques that enable their operation to be "sufficiently transparent to enable interpretation of the system's outputs by humans."

- **Trust and Accountability**: When models are used for critical decisions (e.g., lending, healthcare, criminal justice), stakeholders need to trust the model's decisions and hold it accountable. Treating it as a black box can undermine trust and raise ethical concerns.
- **User Acceptance**: End-users are more likely to accept and adopt models if they can understand the rationale behind the model's decisions, rather than blindly relying on opaque scores.
- **Domain Knowledge Integration**: Incorporating domain expertise and aligning model behaviour with human reasoning can lead to more reliable and trustworthy models, especially in high-stakes domains.
- **Debugging and Improving Models**: Understanding how a model works and its reasoning process can help identify biases, errors, or areas for improvement, ultimately leading to better models.

# Model explainability to gain customer trust

- **Loan application explanations:**
  - A customer applies for a loan but is denied by the bank's ML model.
  - Providing counterfactual explanations, such as "If your annual income were $X higher, and your credit score improved by Y points, you would likely be approved," builds trust by showing transparency and giving actionable feedback.
- **Investment portfolio recommendations:**
  - A wealth management firm uses an ML model to recommend investment portfolios to clients based on their risk profiles and financial goals.
  - Visualizing feature importance and decision boundaries helps explain why certain portfolios were recommended, increasing clients' confidence in the firm's expertise.
- **Fraud detection in transactions:**
  - A bank uses an ML model to flag potentially fraudulent transactions.
  - Providing example-based explanations, such as "This transaction was flagged as suspicious because it shares patterns with other confirmed fraud cases involving X, Y, and Z," helps customers understand the reasoning behind the alert and prevents loss of trust due to false positives.

# Explainability as a reputational risk prevention

**Credit Card Gender Discrimination:** In 2019, multiple reports surfaced that Apple's AI system for determining credit limits for the Apple Card was exhibiting gender discrimination, with lower limits approved for women compared to men with similar credit profiles. The black-box nature of the algorithm made it difficult to pinpoint bias introduction. Had Apple employed interpretability techniques like SHAP to explain each credit limit decision, the drivers of the discrimination may have been more easily auditable and mitigated.

**Money Laundering Detection** In 2018, ING Bank was fined over $900 million for failing to properly monitor customer transactions and catch instances of money laundering. Their automated transaction monitoring system suffered from the "black box" problem, making it difficult to ascertain precisely what factors were triggering or missing certain suspicious activity alerts. More interpretable detection models that could surface the reasoning behind labeling transactions as legitimate or risky could have allowed ING to identify blindspots or biases in their anti-money laundering processes.

**NLP Model Failures** In 2021, JPMorgan Chase reported issues with their AI models for processing incoming emails and communications. The lack of interpretability made it difficult to precisely diagnose why certain models were breaking down, mishandling context in conversations, or exhibiting biased language patterns. This led to disruptions in client communication flows and highlighted the importance of having more transparent NLP systems, especially for sensitive client communications.

# How to facilitate better explainability?

**Use only simpler algorithms as they are generally considered more inherently interpretable than others:**

- **Linear/Logistic Regression:** These models have a simple linear form that makes the relationship between input features and output easy to understand and explain. The coefficients directly represent the feature importances.
- **Decision Trees:** Decision trees split the data based on interpretable rules derived from the features. The tree structure and path to each prediction are human-readable and align with how humans make decisions.
- **Rule-based Classifiers/Scoring:** Rules like "IF condition THEN prediction" directly encode expert knowledge and reasoning in an interpretable way.
- **K-Nearest Neighbours (KNN):** KNN makes predictions by finding the closest examples in the training data, which can provide example-based explanations.

**Even these interpretable models can become opaque as they increase in complexity (e.g., very deep trees, ensembles). And they tend to be less accurate, driving a trade-offs between accuracy and explainability**

# Techniques for enhancing explainability during development

- **Monotonicity constraints:** Enforcing that the model's predictions increase or decrease monotonically with respect to certain input features. This can be desirable when domain knowledge suggests a specific monotonic relationship (e.g., higher income should not decrease the chance of loan approval).
- **Feature sparsity:** Regularizing the model to use only a sparse subset of input features, making the important factors more interpretable.
- **Model distillation:** Training an inherently interpretable model (e.g., decision tree) to mimic the predictions of a more complex black-box model, essentially distilling the knowledge into an interpretable form.
- **Attention regularization:** In deep learning models (e.g. sentiment analysis), regularizing the attention mechanisms to encourage sparse and interpretable attention patterns.
- **Prototype learning:** (e.g. image recognition) Training models to learn prototypical examples that represent different classes or predictions, making the model's decision boundaries more interpretable.

# Techniques for enhancing explainability, post model development, at model level

- **Feature importance methods:** One of the most widely used techniques to understand the overall influence of features in a model is to calculate global feature importance scores. These scores provide a summary of how much each feature contributes to the model's predictions across the entire population or dataset. For ML, a common approach is to use Mean Decrease in Impurity (MDI) or Mean Decrease in Accuracy (MDA) as global feature importance metrics. These metrics calculate the average decrease in the impurity/accuracy of the model when a particular feature is used.
- **Accumulated Local Effects (ALE) plots:** Plot the changes in predictions caused by varying a single feature across the distribution of all other features. They are designed to explicitly handle correlated features and provides a more accurate representation of the feature's effect in the context of the full data distribution than Partial Dependence Plots (PDPs) and Individual Conditional Expectation (ICE) plots but require more computing power.

# Techniques for enhancing explainability, post model development, at case level

- **Example-based Explanations:** Some techniques, like MMD-Critic, can identify the most similar examples from the training data to the case being explained. If the most similar examples in the training data exhibit different patterns for some features (like multiple cash deposits below the threshold)  it suggests that the model may have used those patterns to make the high-risk prediction for the case.
- **Local Interpretable Model-agnostic Explanations (LIME):** LIME works by approximating the model's behaviour locally around the instance being explained (the transaction) using an interpretable model like a linear regression. By examining the coefficients of the local linear model, we can see which features had the highest positive or negative impact on the prediction score, potentially revealing if factors like cash deposit patterns, high-risk entities, or business activity mismatches were influential.
- **SHAP:** computes Shapley values, which attribute each prediction to the contributing features. For a single instance or prediction, it can generate Feature importance values (Indicating how much each feature contributed (positively or negatively) to the specific prediction), visual explanations (such as waterfall plots, showing the cumulative effect of each feature on the prediction)
- **Counterfactual Explanations:** We can generate counterfactual explanations, which show the minimum changes to the transaction features that would result in a different prediction (e.g., low-risk instead of high-risk).

# Future directions and open challenges
# LLM Interpretability

- **Attention Visualization**: Many LLMs, particularly those based on the Transformer architecture, use attention mechanisms to weigh the importance of different input tokens when generating output. By visualizing the attention patterns, we can gain insights into which parts of the input the model focused on when generating a particular response. This can help explain the reasoning behind the model's output.
- **Rationale Generation**: In this approach, the LLM is trained to not only generate the final response but also to provide a rationale or explanation for that response. This can be achieved by fine-tuning the LLM on a dataset that includes both the desired output and the corresponding rationale or explanation.
- **Counterfactual Explanations**: Similar to the techniques used for other machine learning models, they can be generated for LLMs by perturbing the input and observing how the model's output changes. This can help identify the input features or tokens that were most influential in the model's response.
- **Concept Activation Vectors (CAVs):** This technique involves identifying the directions in the model's embedding space that correspond to specific high-level concepts or attributes. By analysing the activation patterns of these concept vectors, we can understand which ones influenced the model's output the most.
- **Confidence Scoring**: LLMs can be trained to produce confidence scores or uncertainty estimates along with their generated responses. These scores can provide insights into how confident the model is about its output, which can be useful for determining when additional explanations or clarifications might be needed.
- **Human-in-the-Loop Explanations**: This can involve having the LLM generate an initial response, which is then reviewed and explained by a human expert in the domain. The human-provided explanations can be used to fine-tune the LLM or to build datasets for training interpretable models.

1.

# Future directions and open challenges: Measuring and evaluating explainability

- **Measuring and evaluating explainability is an active area of research.** There is no universally agreed-upon metric or threshold for what constitutes an acceptable level of explainability, as it is a subjective and context-dependent concept. However, there are some frameworks for assessing the quality and effectiveness of explanations generated by machine learning models.
- **Human evaluation studies:** A common approach where people (e.g., domain experts, end-users) are presented with the explanations generated by different techniques and asked to rate their quality, completeness, and understandability. These evaluations can be based on specific criteria or tasks, such as predicting the model's behaviour based on the explanation or identifying potential biases or errors.
- **Quantitative metrics combination:** capturing different aspects of explainability, such as:
    1. Fidelity: How accurately the explanation reflects the true behaviour of the model.
    2. Consistency: Whether similar inputs produce similar explanations.
    3. Sparsity: How concise and focused the explanation is, highlighting only the most relevant features.
    4. Stability: How robust the explanation is to small perturbations in the input or model parameters.