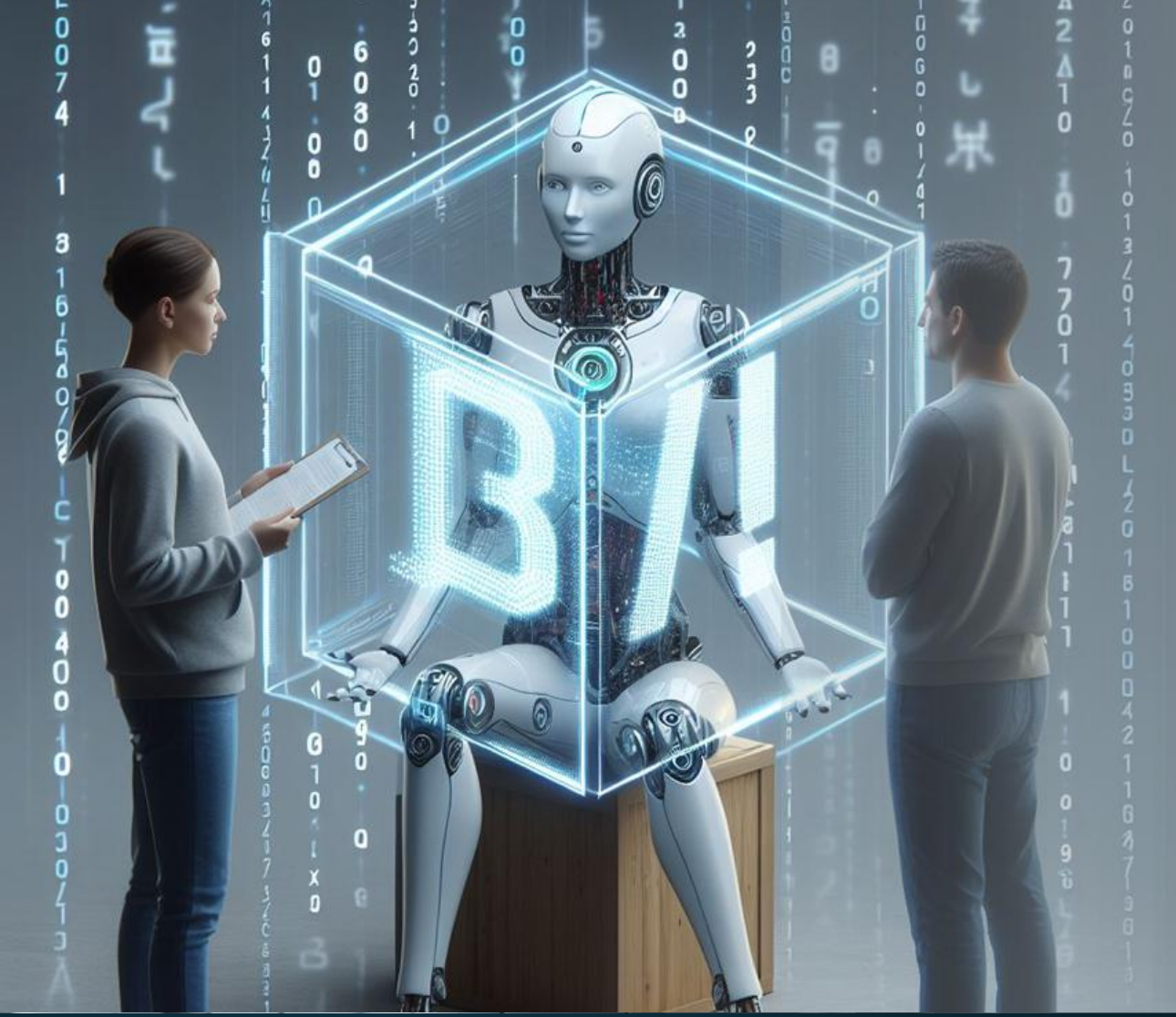
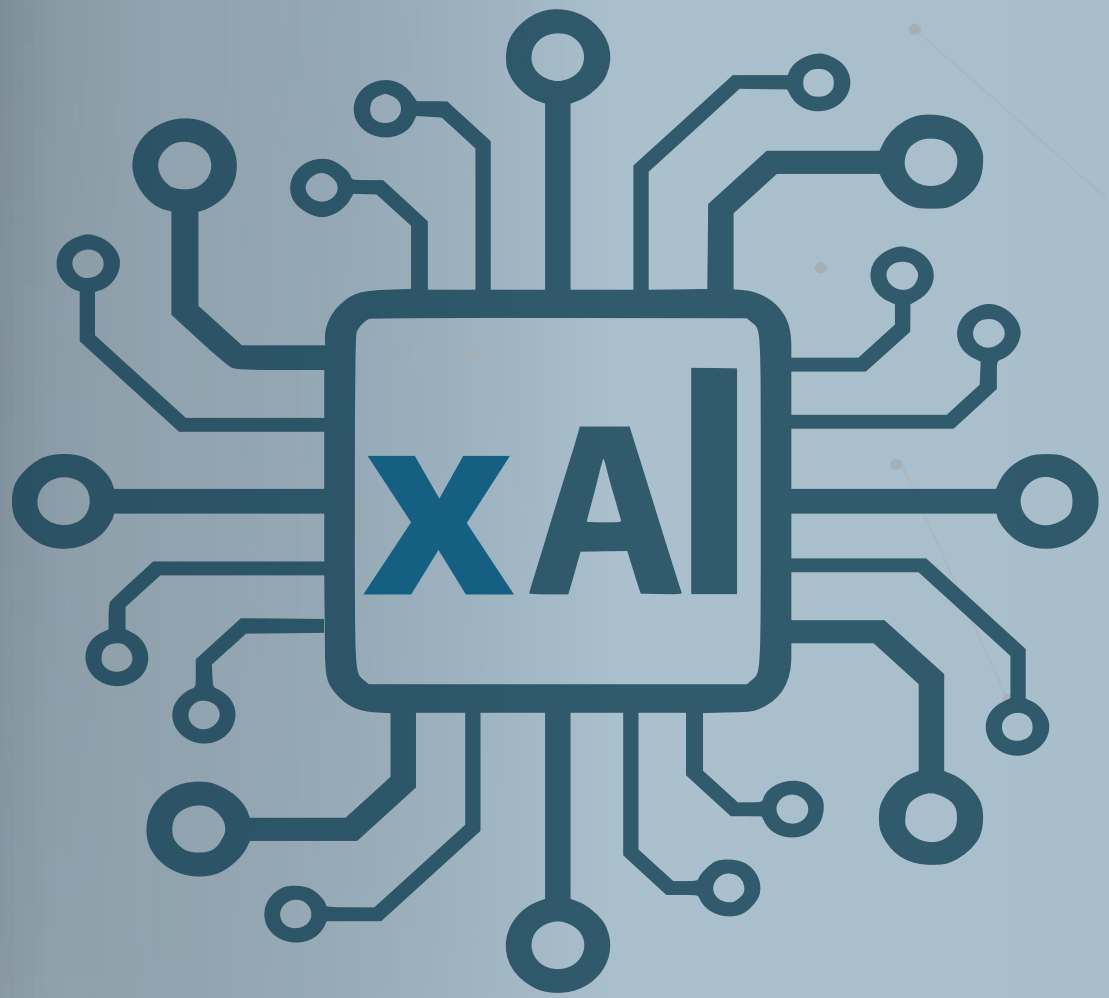




Explainable Artificial Intelligence

Unlocking the Black Box: A Comprehensive Overview of Cutting-Edge Explainable AI Methods

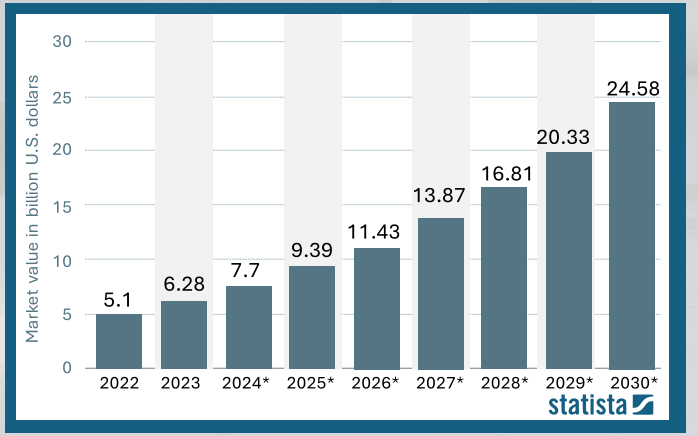
REALM

2025-02-15

EXPLAINABILITY

XAI resembles a black box, and it becomes a challenging task for the experts to understand a logical explanation of the algorithm's decisions.

Explainable AI Market



The Global Explainable AI Market size was valued at USD 5.10 billion in 2022, and it is predicted to reach USD **24.58** billion by 2030, with a CAGR of 21.5% from 2023 to 2030.

CHALLENGES

DATA PRIVACY

AI MODEL COMPLEXITY

HUMAN BIAS ISSUES

USERS UNDERSTANDING

REALM

THE PROBLEM OF EXPLANATION

2025-02-15

AI explanation refers to the process and techniques used to make the decision-making mechanisms of artificial intelligence (AI) models transparent, understandable, and interpretable to humans. This involves clarifying how and why AI systems arrive at specific conclusions or actions, especially in complex models where the decision process is not inherently obvious.

WHAT IS ARTIFICIAL INTELLIGENCE?

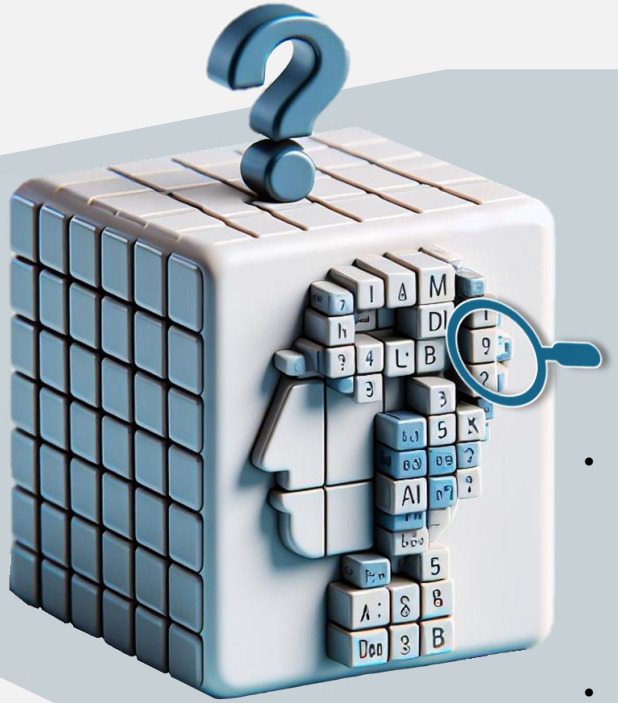
- Artificial Intelligence (AI) encompasses systems or machines that mimic human intelligence to perform tasks and can iteratively improve themselves based on the information they collect.
- AI operates using algorithms, complex mathematical models that make decisions based on input data. These algorithms are powered by weights, which are essentially the parameters that influence decision-making.

THE PROBLEM: INCORRECT WEIGHTS LEAD TO INACCURATE RESULTS

- If a model is trained on biased, insufficient, or irrelevant data, the weights may be adjusted incorrectly. This can lead to the model making erroneous predictions or decisions.
- An example of this issue can be seen in the financial industry, where a loan approval AI system trained on biased historical data might unjustly favor or discriminate against certain groups of applicants. Such a system could deny loans to qualified individuals or approve loans for those who are not likely to repay, based on flawed weight adjustments.

THE SOLUTION: EXPLAINING AI THROUGH RULES OR DECISION TREES:

The XAI's approach is to translate the model's complex decisions into understandable rules or decision trees.



How Do Weights Work in AI?

- Weights in AI models are determined during the training process, where the model learns from a dataset to make accurate predictions or decisions.
- These weights adjust as the model learns, striving to minimize errors in its predictions compared to the actual outcomes.

OVERVIEW OF XAI METHODS

In the quest to make artificial intelligence (AI) more interpretable and transparent, Explainable AI (XAI) employs a variety of methods and tools designed to shed light on the decision-making processes of complex models. These methods can broadly be categorized into model-agnostic, model-specific, and visualization-based approaches, each offering unique insights into how AI systems operate.

Model-Agnostic Methods

These methods are designed to work with any machine learning model, providing flexibility and broad applicability. They do not require access to the model's internal architecture, making them highly versatile for explaining different types of AI models.

Model-Specific Approaches

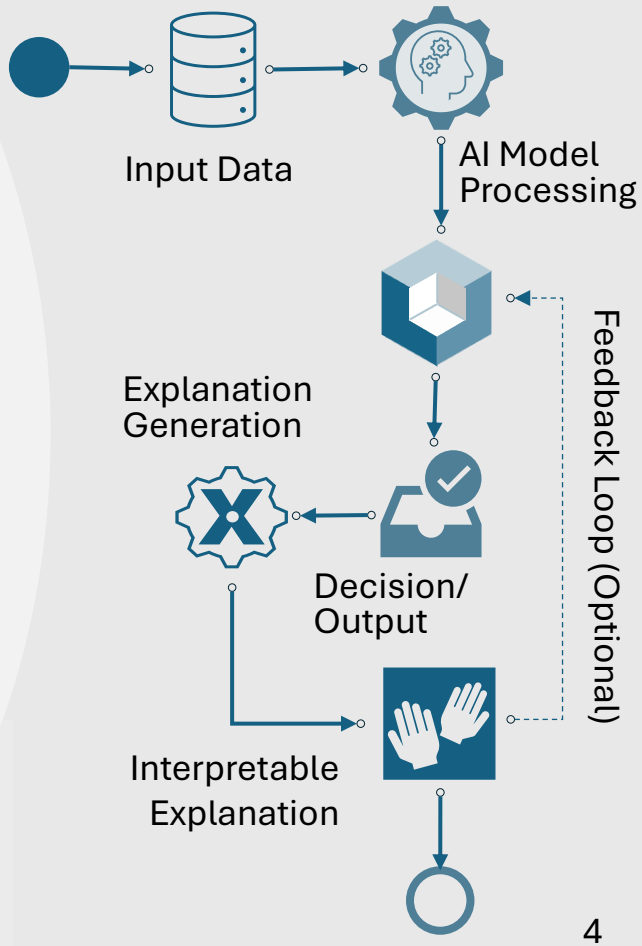
In contrast, model-specific methods are tailored to specific types of models. They leverage the internal mechanics of the models they are designed to explain, offering deeper insights into the model's decision-making process. However, their applicability is limited to certain model types

Visualization-Based Approaches

Visualization tools and techniques play a crucial role in XAI by providing intuitive and accessible explanations. They transform complex model outputs into visual formats that are easier to understand, making them an invaluable resource for both technical and non-technical stakeholders.

POPULAR XAI METHODS			
LIME (Local Interpretable Model-Agnostic Explanations)	Generates local explanations by using simple models to approximate the predictions of the AI model in the vicinity of a particular point.	LRP (Layer-wise Relevance Propagation)	Assigns relevance scores to input features, reflecting their contribution to the model's prediction.
SHAP (SHapley Additive Explanations)	uses Shapley's game to compute the contribution of each input element to the model's prediction	GAM (Generalized Additive Model)	Represents the AI model as a sum of simple functions, each of which explains the dependence of the prediction on one of the input parameters.
Counterfactual Explanations	Generates examples of input data that lead to a change in the prediction of the model.	Rationalization	Rationalization uses inference techniques to create rule-based explanations.

XAI Process



UNLOCKING THE FUTURE OF AI WITH ATTRIBUTION METHODS

2025-02-15

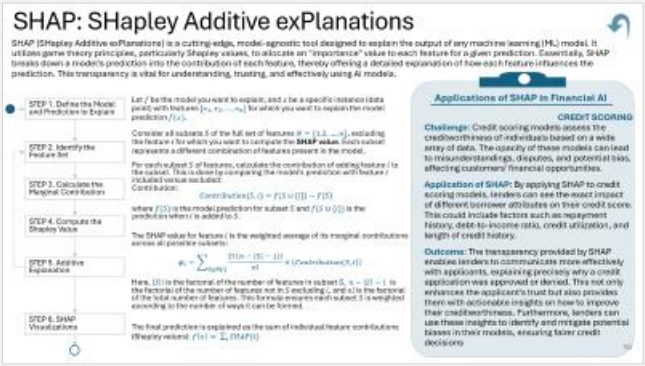
Among the various approaches to XAI, attribution methods like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have emerged as key players. These methods offer a nuanced view into the "why" behind AI's decisions, presenting a weighted significance of the features that drive model outcomes.

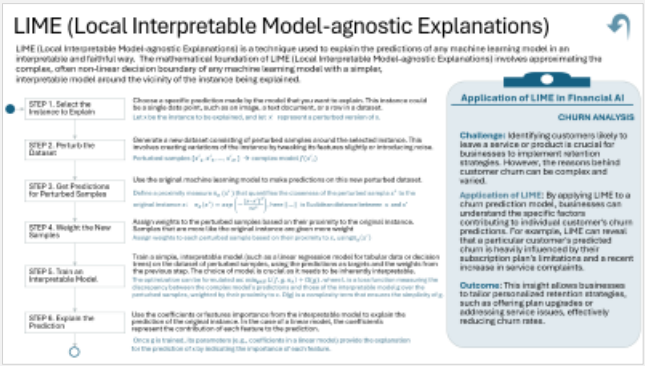
Redefining Interpretability: Attribution methods are revolutionizing how we interpret AI models, shifting the paradigm from opaque "black boxes" to transparent systems where decision paths are not just visible but quantifiable.
Application Examples: In finance, SHAP has elucidated credit scoring decisions, while LIME has shed light on healthcare AI, providing clear reasoning behind patient diagnosis predictions.

The Power of Feature Attribution: By assigning a quantitative value to each feature's contribution to a prediction, these methods demystify the AI decision-making process, empowering users with actionable insights.
Application Examples: Marketing analytics utilize attribution methods to pinpoint which customer interactions most influence purchase decisions, enhancing targeted strategies.

Facilitating Trust and Ethical AI: Attribution methods play a critical role in building trust between AI systems and their human users by ensuring decisions are fair, understandable, and aligned with ethical standards.
Application Examples: In law enforcement, LIME and SHAP have been applied to predictive policing models to ensure transparency and prevent bias in crime prediction algorithms.

Promoting Model Improvement and Innovation: Beyond explanation, attribution methods offer a feedback loop for model refinement, highlighting inaccuracies or biases in feature weighting that can be corrected for more accurate future predictions.
Application Examples: In autonomous vehicle development, attribution methods help identify and correct sensor input weights, improving decision accuracy in dynamic driving environments.





MARKETING ATTRIBUTION

A LAYERED APPROACH TO XAI ARCHITECTURE

A comprehensive, layered approach to XAI architecture not only ensures clarity and transparency in AI's decision-making processes but also enhances model reliability and stakeholder trust. This slide delves into the necessity of adopting a complex, layered framework to effectively unpack the "black box" of AI, providing meaningful insights and explanations.

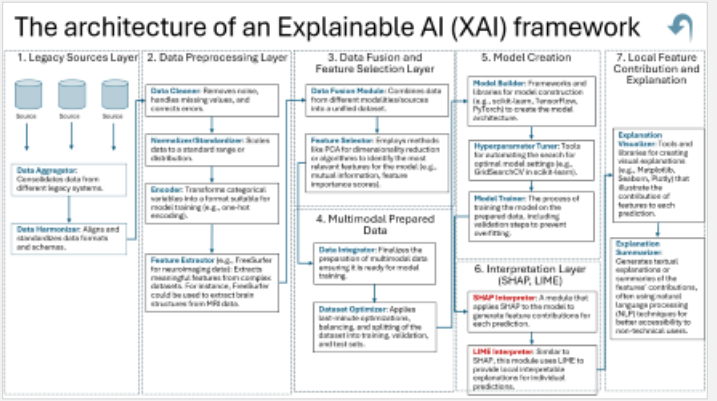
Complexity Demands Structure. As AI systems grow in complexity, a structured, layered approach is paramount for dissecting and understanding their inner workings. Each layer focuses on a specific aspect of the AI pipeline, from data sourcing to model explanation, ensuring thorough analysis and clarity.

Holistic Data Handling. Effective XAI begins with comprehensive data management, from aggregation of legacy sources to advanced preprocessing and fusion. Each stage is crucial for ensuring the quality and integrity of data feeding into AI models.

Interpretable Model Development. Creating interpretable models requires careful construction and training, with an emphasis on selecting algorithms that balance predictive power with explainability.
Implication: Through deliberate model selection and hyperparameter tuning within its dedicated layer, AI developers can enhance both model performance and the feasibility of subsequent explanations.

Advanced Explanation Techniques. Employing advanced attribution methods like SHAP and LIME in a distinct interpretation layer allows for detailed insights into how AI models arrive at their decisions, illuminating the contribution of individual features.

Ensuring Model and Data Security. Incorporating security and privacy considerations into each layer is essential for protecting sensitive information and complying with regulatory standards.



IMPLICATIONS:

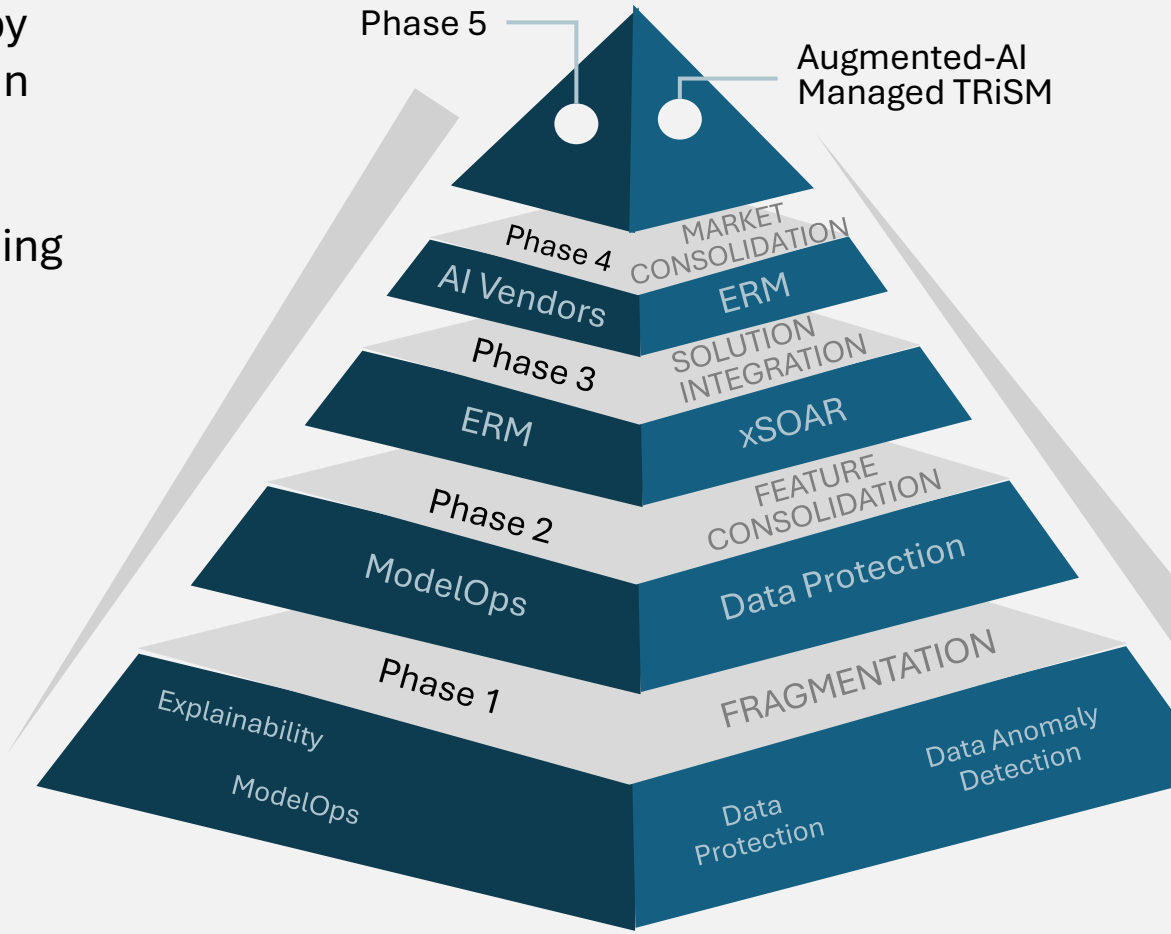
- By meticulously preparing and optimizing data across multiple layers, AI systems can make more accurate and interpretable predictions, reducing biases and errors.
- This dedicated layer for model interpretation provides stakeholders with clear, actionable explanations, fostering trust and facilitating more informed decision-making.
- A layered architecture with built-in security measures safeguards against data breaches and unauthorized access, ensuring ethical AI utilization.

AI TrISM : Building Trust through Transparency and Security

2025-02-15

In the evolving landscape of artificial intelligence, ensuring the trustworthiness, security, and ethical deployment of AI systems is paramount. AI Trust, Risk, and Security Management (TrISM) emerges as a critical framework in this endeavor, seamlessly aligning with Explainable AI (XAI) principles to foster transparent, secure, and reliable AI solutions. This slide explores the convergence of AI TrISM and XAI, highlighting their collaborative role in enhancing trust in AI technologies.

- **AI TrISM** and **XAI** together form the cornerstone of trust in AI by emphasizing transparency, explainability, and accountability in AI systems.
- **XAI** contributes to **TrISM** by enabling a deeper understanding of AI models, which is essential for identifying, assessing, and mitigating risks.
- A thorough grasp of AI model behaviors, facilitated by **XAI**, is crucial for implementing effective security measures and protecting against malicious exploits.
- The convergence of **AI TrISM** and **XAI** is instrumental in ensuring AI systems operate within ethical boundaries and comply with existing and emerging regulations.

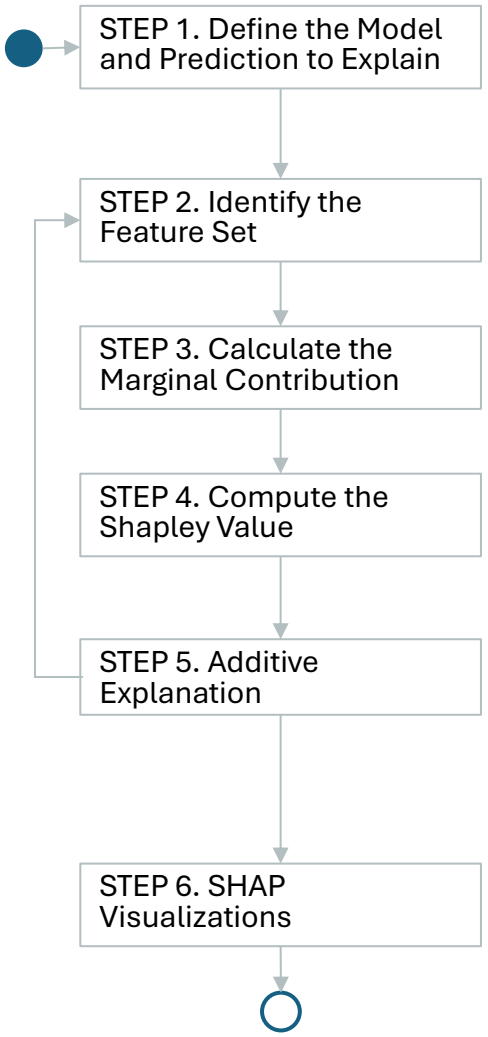


THANK YOU

APPENDIX

SHAP: SHapley Additive exPlanations

SHAP (SHapley Additive exPlanations) is a cutting-edge, model-agnostic tool designed to explain the output of any machine learning (ML) model. It utilizes game theory principles, particularly Shapley values, to allocate an "importance" value to each feature for a given prediction. Essentially, SHAP breaks down a model's prediction into the contribution of each feature, thereby offering a detailed explanation of how each feature influences the prediction. This transparency is vital for understanding, trusting, and effectively using AI models.



Let f be the model you want to explain, and x be a specific instance (data point) with features $[x_1, x_2, \dots, x_n]$ for which you want to explain the model prediction $f(x)$.

Consider all subsets S of the full set of features $N = \{1,2, \dots, n\}$, excluding the feature i for which you want to compute the **SHAP value**. Each subset represents a different combination of features present in the model.

For each subset S of features, calculate the contribution of adding feature i to the subset. This is done by comparing the model's prediction with feature i included versus excluded:
Contribution:

$$Contribution(S, i) = f(S \cup \{i\}) - f(S)$$

where $f(S)$ is the model prediction for subset S and $f(S \cup \{i\})$ is the prediction when i is added to S.

The SHAP value for feature i is the weighted average of its marginal contributions across all possible subsets:

$$\phi_i = \sum_{S \subseteq N\{i\}} \frac{S!\,(n - |S| - 1)!}{n!} \times (Contribution(S, i))$$

Here, $|S|!$ is the factorial of the number of features in subset S, $n - |S| - 1$ is the factorial of the number of features not in S excluding i, and $n!$ is the factorial of the total number of features. This formula ensures each subset S is weighted according to the number of ways it can be formed.

The final prediction is explained as the sum of individual feature contributions (Shapley values): $f(x) = \sum_i SHAP(i)$

Applications of SHAP in Financial AI

CREDIT SCORING

Challenge: Credit scoring models assess the creditworthiness of individuals based on a wide array of data. The opacity of these models can lead to misunderstandings, disputes, and potential bias, affecting customers' financial opportunities.

Application of SHAP: By applying SHAP to credit scoring models, lenders can see the exact impact of different borrower attributes on their credit score. This could include factors such as repayment history, debt-to-income ratio, credit utilization, and length of credit history.

Outcome: The transparency provided by SHAP enables lenders to communicate more effectively with applicants, explaining precisely why a credit application was approved or denied. This not only enhances the applicant's trust but also provides them with actionable insights on how to improve their creditworthiness. Furthermore, lenders can use these insights to identify and mitigate potential biases in their models, ensuring fairer credit decisions

LIME (Local Interpretable Model-agnostic Explanations)

LIME (Local Interpretable Model-agnostic Explanations) is a technique used to explain the predictions of any machine learning model in an interpretable and faithful way. The mathematical foundation of LIME (Local Interpretable Model-agnostic Explanations) involves approximating the complex, often non-linear decision boundary of any machine learning model with a simpler, interpretable model around the vicinity of the instance being explained.

STEP 1. Select the Instance to Explain

Choose a specific prediction made by the model that you want to explain. This instance could be a single data point, such as an image, a text document, or a row in a dataset.

Let x be the instance to be explained, and let x' represent a perturbed version of x.

STEP 2. Perturb the Dataset

Generate a new dataset consisting of perturbed samples around the selected instance. This involves creating variations of the instance by tweaking its features slightly or introducing noise.

Perturbed samples $\{x'_1, x'_2, \ldots, x'_N\}$ → complex model $f(x'_i)$

STEP 3. Get Predictions for Perturbed Samples

Use the original machine learning model to make predictions on this new perturbed dataset.

Define a proximity measure $\pi_x(x')$ that quantifies the closeness of the perturbed sample x' to the original instance x: $\pi_x(x') = exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$, here $\|\ldots\|$ is Euclidean distance between x and x'

STEP 4. Weight the New Samples

Assign weights to the perturbed samples based on their proximity to the original instance. Samples that are more like the original instance are given more weight

Assign weights to each perturbed sample based on their proximity to x, using $\pi_x(x')$

STEP 5. Train an Interpretable Model

Train a simple, interpretable model (such as a linear regression model for tabular data or decision trees) on the dataset of perturbed samples, using the predictions as targets and the weights from the previous step. The choice of model is crucial as it needs to be inherently interpretable.

The optimization can be formulated as: $\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$, where L is a loss function measuring the discrepancy between the complex model's predictions and those of the interpretable model g over the perturbed samples, weighted by their proximity to x. Ω(g) is a complexity term that ensures the simplicity of g.

STEP 6. Explain the Prediction

Use the coefficients or features importance from the interpretable model to explain the prediction of the original instance. In the case of a linear model, the coefficients represent the contribution of each feature to the prediction.

Once g is trained, its parameters (e.g., coefficients in a linear model) provide the explanation for the prediction of x by indicating the importance of each feature.

Application of LIME in Financial AI

CHURN ANALYSIS

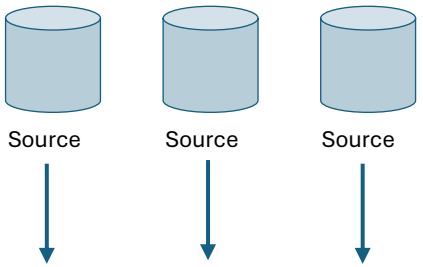
Challenge: Identifying customers likely to leave a service or product is crucial for businesses to implement retention strategies. However, the reasons behind customer churn can be complex and varied.

Application of LIME: By applying LIME to a churn prediction model, businesses can understand the specific factors contributing to individual customer's churn predictions. For example, LIME can reveal that a particular customer's predicted churn is heavily influenced by their subscription plan's limitations and a recent increase in service complaints.

Outcome: This insight allows businesses to tailor personalized retention strategies, such as offering plan upgrades or addressing service issues, effectively reducing churn rates.

The architecture of an Explainable AI (XAI) framework

1. Legacy Sources Layer



Data Aggregator: Consolidates data from different legacy systems.

Data Harmonizer: Aligns and standardizes data formats and schemas.

2. Data Preprocessing Layer

Data Cleaner: Removes noise, handles missing values, and corrects errors.

Normalizer/Standardizer: Scales data to a standard range or distribution.

Encoder: Transforms categorical variables into a format suitable for model training (e.g., one-hot encoding).

Feature Extractor (e.g., FreeSurfer for neuroimaging data): Extracts meaningful features from complex datasets. For instance, FreeSurfer could be used to extract brain structures from MRI data.

3. Data Fusion and Feature Selection Layer

Data Fusion Module: Combines data from different modalities/sources into a unified dataset.

Feature Selector: Employs methods like PCA for dimensionality reduction or algorithms to identify the most relevant features for the model (e.g., mutual information, feature importance scores).

4. Multimodal Prepared Data

Data Integrator: Finalizes the preparation of multimodal data ensuring it is ready for model training.

Dataset Optimizer: Applies last-minute optimizations, balancing, and splitting of the dataset into training, validation, and test sets.

5. Model Creation

Model Builder: Frameworks and libraries for model construction (e.g., scikit-learn, TensorFlow, PyTorch) to create the model architecture.

Hyperparameter Tuner: Tools for automating the search for optimal model settings (e.g., GridSearchCV in scikit-learn).

Model Trainer: The process of training the model on the prepared data, including validation steps to prevent overfitting.

6. Interpretation Layer (SHAP, LIME)

SHAP Interpreter: A module that applies SHAP to the model to generate feature contributions for each prediction.

LIME Interpreter: Similar to SHAP, this module uses LIME to provide local interpretable explanations for individual predictions.

7. Local Feature Contribution and Explanation

Explanation Visualizer: Tools and libraries for creating visual explanations (e.g., Matplotlib, Seaborn, Plotly) that illustrate the contribution of features to each prediction.

Explanation Summarizer: Generates textual explanations or summaries of the features' contributions, often using natural language processing (NLP) techniques for better accessibility to non-technical users.