Training Course on Explainable AI (XAI) and Model Interpretability

Professional Development Training Course Outline

Category	Data Science	**Duration**	5 Days
Base Fee	$1,500 USD	**Delivery Mode**	Online / On-site Hybrid

Course Introduction

The proliferation of Artificial Intelligence (AI) and Machine Learning (ML) models across critical sectors like healthcare, finance, and autonomous systems has underscored an urgent need for transparency and trust. While "black box" models often deliver superior predictive performance, their opaque decision-making processes pose significant challenges for accountability, debugging, and regulatory compliance. Training Course on Explainable AI (XAI) & Model Interpretability addresses this critical gap, equipping participants with the essential techniques to demystify complex AI systems, fostering greater confidence and responsible AI deployment.

This course dives deep into **state-of-the-art model-agnostic explanation methods**, specifically focusing on **LIME (Local Interpretable Model-agnostic Explanations)** and **SHAP (SHapley Additive exPlanations)**. Through hands-on exercises and real-world case studies, learners will master the practical application of these powerful tools, enabling them to generate actionable insights from intricate ML models. By understanding *why* an AI system makes a particular prediction, professionals can build more robust, ethical, and trustworthy AI solutions that align with evolving **Responsible AI** principles and **AI governance** frameworks.

Programme Curriculum

Training Course on Explainable AI (XAI) & Model Interpretability: Techniques for Understanding 'Black Box' Models (LIME, SHAP)

Introduction

The proliferation of Artificial Intelligence (AI) and Machine Learning (ML) models across critical sectors like healthcare, finance, and autonomous systems has underscored an urgent need for transparency and trust. While "black box" models often deliver superior predictive performance, their opaque decision-making processes pose significant challenges for accountability, debugging,

and regulatory compliance. Training Course on Explainable AI (XAI) & Model Interpretability addresses this critical gap, equipping participants with the essential techniques to demystify complex AI systems, fostering greater confidence and responsible AI deployment.

This course dives deep into **state-of-the-art model-agnostic explanation methods**, specifically focusing on **LIME (Local Interpretable Model-agnostic Explanations)** and **SHAP (SHapley Additive exPlanations)**. Through hands-on exercises and real-world case studies, learners will master the practical application of these powerful tools, enabling them to generate actionable insights from intricate ML models. By understanding *why* an AI system makes a particular prediction, professionals can build more robust, ethical, and trustworthy AI solutions that align with evolving **Responsible AI** principles and **AI governance** frameworks.

Course Duration

5 days

Course Objectives

1. Understand the inherent opacity of complex machine learning models and the imperative for **Explainable AI (XAI)**.
2. Clearly define and distinguish between model interpretability, explainability, and transparency in the AI lifecycle.
3. Gain proficiency in applying **LIME** and **SHAP** as leading model-agnostic explanation frameworks.
4. Apply LIME to generate **local explanations** for individual predictions, understanding feature importance at the instance level.
5. Leverage SHAP to derive both **local and global feature attributions**, providing a unified understanding of model behavior.
6. Accurately quantify the **impact of individual features** on model predictions using XAI techniques.
7. Employ interpretability methods to identify and address **algorithmic bias** and fairness concerns within AI models.
8. Learn to assess the quality and reliability of explanations using relevant **XAI evaluation metrics**.
9. Understand how to incorporate XAI techniques into **MLOps pipelines** for continuous monitoring and improvement of AI systems.
10. Develop skills to communicate complex AI explanations to diverse stakeholders, fostering **AI literacy** and trust.
11. Understand the role of XAI in meeting **AI regulations** and fostering **responsible AI development**.
12. Gain an overview of emerging XAI techniques such as **counterfactual explanations** and **causal inference** in AI.
13. Solve practical business problems by applying XAI to various domains, including **healthcare AI**, **financial services AI**, and **autonomous systems**.

Organizational Benefits

- Build greater confidence in AI systems among users, stakeholders, and regulatory bodies, leading to wider adoption and utilization.
- Expedite the identification and resolution of model errors, biases, and performance degradation, leading to more robust and accurate AI.

- Ensure compliance with evolving data privacy and AI ethics regulations (e.g., GDPR, AI Act), mitigating legal and reputational risks associated with opaque AI.
- Empower business users and domain experts to make more informed decisions by understanding the rationale behind AI predictions.
- Encourage the development of more complex and impactful AI solutions by providing the tools to understand and control their behavior.
- Direct resources more effectively by understanding which features truly drive model outcomes, leading to more efficient data collection and model training.
- Position the organization at the forefront of responsible and ethical AI implementation, a growing differentiator in the market.

Target Audience

1. Data Scientists.
2. Machine Learning Engineers
3. AI Product Managers
4. AI/ML Researchers.
5. Business Analysts & Domain Experts.
6. Regulatory & Compliance Officers.
7. Senior Management & Executives.
8. Anyone interested in AI Accountability.

Course Outline

Module 1: Introduction to Explainable AI (XAI) & The Black Box Problem

- Understanding the AI Black Box: Why complex models (e.g., deep neural networks, ensemble methods) are often opaque.
- The Imperative for XAI: Discussing the need for transparency, trust, and accountability in AI.
- Key Concepts: Differentiating between interpretability, explainability, and transparency.
- Types of XAI: Overview of intrinsic (white box) vs. post-hoc (black box) explainability.
- Challenges in XAI: Exploring limitations and trade-offs in explainability.
- **Case Study**: *Financial Lending Decisions*. Analyzing a scenario where a loan application is rejected by an opaque AI model, highlighting the legal and ethical challenges of a "black box" decision.

Module 2: Foundations of Model Interpretability

- Simple Interpretable Models: Revisiting linear regression, logistic regression, and decision trees as inherently interpretable models.
- Feature Importance & Permutation Importance: Understanding global feature relevance.
- Partial Dependence Plots (PDPs): Visualizing the marginal effect of features on model predictions.
- Individual Conditional Expectation (ICE) Plots: Examining individual predictions' dependency on features.
- Surrogate Models: Concepts of global and local surrogate models for approximation.
- **Case Study**: *Predicting Customer Churn*. Using PDPs and ICE plots to understand which customer attributes globally and individually contribute to churn predictions for a telecom company.

Module 3: Local Interpretable Model-agnostic Explanations (LIME)

- LIME Principles: How LIME approximates a black-box model locally with an interpretable surrogate model.
- Perturbation & Sampling: Understanding how LIME generates perturbed data points.
- Weighted Linear Models: The role of local weighting and simple model fitting.
- Interpreting LIME Explanations: Visualizing and understanding feature contributions for single predictions.
- Practical Implementation with LIME: Hands-on exercises using the lime library in Python.
- **Case Study**: *Medical Diagnosis*. Explaining a specific patient's diagnosis by an AI model (e.g., tumor detection) using LIME, pinpointing which image regions were most influential for that particular prediction.

Module 4: SHapley Additive exPlanations (SHAP)

- Game Theory & Shapley Values: Introducing the theoretical foundation of SHAP.
- Unified Feature Attribution: How SHAP provides a consistent and theoretically sound measure of feature importance.
- KernelSHAP & TreeSHAP: Exploring different SHAP algorithms for various model types.
- Interpreting SHAP Plots: Force plots, summary plots, and dependence plots for holistic analysis.
- Practical Implementation with SHAP: Hands-on exercises using the shap library in Python.
- **Case Study**: *Fraud Detection in Banking*. Analyzing a flagged transaction using SHAP to determine the exact contribution of various transaction parameters (amount, location, frequency) to the fraud prediction.

Module 5: Comparing LIME and SHAP & Choosing the Right Technique

- Strengths and Weaknesses: A comparative analysis of LIME and SHAP.
- When to Use Which: Practical guidelines for selecting the appropriate XAI technique based on use case and model complexity.
- Consistency & Stability: Addressing potential issues with explanation stability.
- Computational Considerations: Performance implications of different XAI methods.
- Combining XAI Techniques: Strategies for leveraging multiple explanation methods for deeper insights.
- **Case Study**: *Credit Score Prediction*. Applying both LIME and SHAP to the same credit score prediction and comparing the insights gained, discussing which method offers more valuable information for different stakeholders (e.g., loan applicant vs. risk analyst).

Module 6: Evaluating Explainability & Ensuring Robustness

- Metrics for Explanation Quality: Faithfulness, stability, comprehensibility, and informativeness.
- Human-in-the-Loop Evaluation: Incorporating human feedback for validating explanations.
- Robustness of Explanations: Assessing how sensitive explanations are to small input perturbations.
- Adversarial Explanations: Understanding how explanations can be manipulated.
- Best Practices for Deploying XAI: Guidelines for responsible and effective XAI implementation.
- **Case Study**: *Autonomous Vehicle Safety*. Evaluating the robustness of an XAI system that explains an autonomous car's decision to brake, ensuring the explanation remains consistent and reliable under varying road conditions

Module 7: Ethical Considerations & Responsible AI

- Algorithmic Bias Detection: Using XAI to uncover and quantify biases in model decisions.
- Fairness in AI: Discussing different notions of fairness and how XAI supports ethical AI.
- Privacy Concerns in XAI: Balancing transparency with data privacy.
- Regulatory Landscape: Overview of emerging AI regulations and their implications for explainability.
- Building Trustworthy AI Systems: Integrating XAI into a broader Responsible AI framework.
- **Case Study**: *Recidivism Prediction*. Examining an AI model used in the justice system to predict recidivism and using XAI to detect and illustrate potential biases against certain demographic groups, leading to discussions on fairness mitigation strategies

Module 8: Advanced XAI Topics & Future Trends

- Counterfactual Explanations: Generating "what if" scenarios to understand decision changes.
- Causal Inference & XAI: Moving beyond correlation to understand causal relationships.
- Explainability for Specific AI Domains: Deep learning, natural language processing (NLP), and computer vision.
- XAI in MLOps & Monitoring: Continuous explainability in production environments.
- Emerging XAI Techniques & Research Directions: Latest advancements and future outlook for the field.
- **Case Study**: *Personalized Marketing Campaigns*. Using counterfactual explanations to show a marketing manager what changes in a customer's profile would lead to a different product recommendation from an AI system.

Training Methodology

This course employs a participatory and hands-on approach to ensure practical learning, including:

- Interactive lectures and presentations.
- Group discussions and brainstorming sessions.
- Hands-on exercises using real-world datasets.
- Role-playing and scenario-based simulations.
- Analysis of case studies to bridge theory and practice.
- Peer-to-peer learning and networking.
- Expert-led Q&A sessions.
- Continuous feedback and personalized guidance.

Register as a group from 3 participants for a Discount

Send us an email: info@fineskilltrainingcenter.org or call *+254769199797*

Certification

Upon successful completion of this training, participants will be issued with a globally- recognized certificate.

Tailor-Made Course

We also offer tailor-made courses based on your needs.

Key Notes

a. The participant must be conversant with English.

b. Upon completion of training the participant will be issued with an Authorized Training Certificate

c. Course duration is flexible and the contents can be modified to fit any number of days.

d. The course fee includes facilitation training materials, 2 coffee breaks, buffet lunch and A Certificate upon successful completion of Training.

e. One-year post-training support Consultation and Coaching provided after the course.

f. Payment should be done at least a week before commence of the training, to Fineskill Training Center account, as indicated in the invoice so as to enable us prepare better for you.

Upcoming Scheduled Sessions

Start Date	End Date	Location	Price
23 Nov 2026	27 Nov 2026	Online	$1,000
23 Nov 2026	27 Nov 2026	Physical	$1,500
23 Nov 2026	27 Nov 2026	Physical	$3,000
23 Nov 2026	27 Nov 2026	Physical	$3,000
23 Nov 2026	27 Nov 2026	Physical	$6,000
23 Nov 2026	27 Nov 2026	Physical	$4,500
23 Nov 2026	27 Nov 2026	Physical	$6,000
23 Nov 2026	27 Nov 2026	Physical	$6,000

Ready to enroll or request a customized program? Book online or contact us:

Register Online Now

Email: info@fineskilltrainingcenter.com | Website: www.fineskilltrainingcenter.com