Text, logo

Description automatically generatedLogo

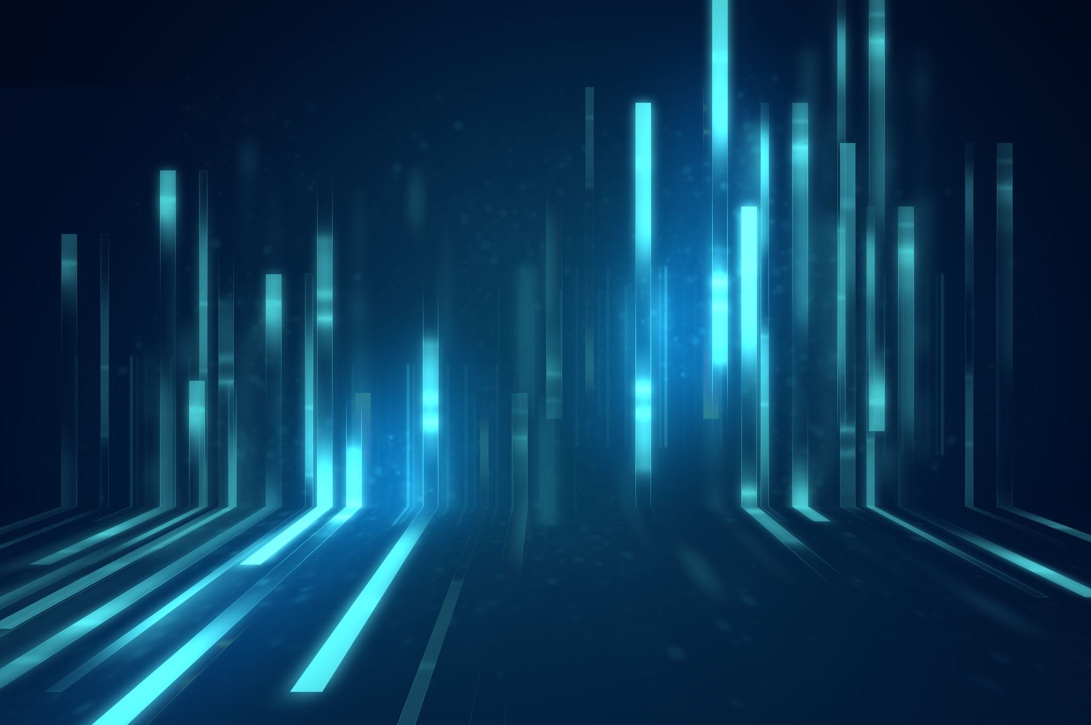
Description automatically generated with medium confidence

­­­

Transparency Intermediate

Module 3

**Applying Technological Concepts**



­­­

# Introduction

In this module we discuss how to approach transparency and explainability alongside technical development:

* Assessing transparency and explainability requirements of the problem at hand
* Discuss ways to perform a detailed technical breakdown of model performance and algorithmic development that assesses the performance, transparency, justifiability and accountability of an algorithms
* Provide an overview of quantitative metrics and indicators available (e.g., feature importance analysis, data provenance, SHAP values)
* Analyse issues of transparency and explainability in the police and security domain through concrete examples and exercises drawn from CESIUM product, which includes a custom user-facing xAI component

# Reading

* Machine Learning Metrics – further reading

<https://www.altexsoft.com/blog/machine-learning-metrics>

* Interpretable Machine Learning

<https://christophm.github.io/interpretable-ml-book/simple.html>

(chapter 5, interpretable models)

This will give an overview of interpretable models (white-box) and their desirable interpretable features

* Model-Agnostic Explanation

For some good insight on desirable features of model-agnostic explanation, read here – see chapter 6

<https://christophm.github.io/interpretable-ml-book/agnostic.html>

* The xAI Pyramid

DALEX v 1.0 and the Explanatory Model Analysis

<https://medium.com/@ModelOriented/dalex-v-1-0-and-the-explanatory-model-analysis-419585a4ba91>

(a more thorough explanation of the xAI pyramid)

­­­



**Self-assessment pass/fail questions**

1. Which of the following regarding explainability in ML (XAI) is NOT true?

a. XAI deals with a set of techniques that make decision making in ML more transparent

b. Justification by showing a machine learning model's accuracy will satisfy the end user

c. Explaining AI with an ethical approach will help users trust AI

d. XAI is mandatory in many countries

e. All of these are true

2. Which of the following models are classified as transparent/white box models?

a. Decision Trees

b. Logistic Regression

c. Linear Regression

d. K-nearest neighbours

e. All of the above

3. Which of the following is NOT correct about local explanations?

a. Collections of local explanations can be used for global explainability

b. Local explanations help to determine if individual predictions are made for the right reasons

c. Local explanations answer questions as to which variables contribute more to the prediction for instances of interest

d. Individual local explanations help to investigate the variation of model performance for different sub-groups of the data

e. All of these are true

4. Why is explainability important in ML?

a. Understanding model behaviour

b. Addressing associated biases with the model

c. Engaging various stakeholders as part of the project lifecycle

d. Building trust and confidence with the end user

e. Identifying limitations and mechanisms to improve model generalization

f. All of the above

5. Which of the following feature importance methods DO NOT allow global explainability?

a. Permutation Feature importance (PFI)

b. Partial dependence plots (PDPs)

c. Feature importance via a model intrinsic model parameter.

d. Individual conditional expectations (ICEs)

e. Accumulated Local Effects (ALE)

6. Which of the following about SHAP is not correct?

a. Allows both local and global explainability

b. Summary plots in SHAP allow global explainability

c. SHAP works only with Decision Trees and Neural networks

d. SHAP is a unified framework of methods and explainers that approximate shapely values

e. Tree SHAP works with tree-based machine learning models

7. Which of the following describes target/concept drift in machine learning model monitoring?

a. Target drift investigates change in predictive power of a model

b. Target drift can occur due to model unlearning patterns it had learned before due to changes in the environment

c. Kolmogorov-Smirnov and Chi-squared test can be used to measure target drift

d. Target drift helps to investigate how feature correlations change with the target variable

e. All of the above

8. Which of the following classification metrics should not be used with imbalanced datasets?

a. Precision

b. Recall

c. F1

d. Accuracy

e. Balanced Accuracy

9. Which the following is a fairness metrics in the fairlearn library apply to regression-based ML tasks?

a. Bounded group loss

b. Equalized Odds

c. Demographic Parity

d. True Positive Rate Parity

e. False Positive Rate Parity

10. Which of the following are useful techniques to gain an understanding of your data?

a. Identify outliers in dataset

b. Look for class imbalances in the dataset

c. Investigate data distributions via data profiling

d. Understand feature correlations with the target variable via correlation matrices

e. All of the above

11. Which of the following considerations when developing an end-to-end machine learning solution?

a. Who are the end users? What insight is needed?

b. What data do we have available? Is the quality of the data suitable to do machine learning?

c. How is the end user currently tacking the problem?

d. Choice of model to use and how to make it explainable to the end user

e. All of the above

12. Which of the following bias refers to the training dataset not being representative and diverse for all cohorts of population?

a. Selection bias

b. Confirmation bias

c. Reporting bias

d. Label bias

e. All of the above

**Answers**

Qs 1) b, 2) e, 3) d, 4) f, 5) d, 6) c, 7) e, 8) d, 9) a, 10) e, 11) e, 12) a Qs

# References

Masís, S. (2021). Interpretable Machine Learning with Python: Learn to build interpretable high-performance models with hands-on real-world examples. Packt Publishing Ltd.

Molnar, C. (2019). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018

Lundberg, S., Allen, P. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. [online] Available at: <https://arxiv.org/pdf/1705.07874.pdf>

Shapley, L.S. (1952). A Value for N-Person Games. [online] www.rand.org. Available at: <https://www.rand.org/pubs/papers/P295.html> [Accessed 22 Sep. 2022]

Wu, T., Ribeiro, M., Heer, J. and Weld, D. (2021). Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. [online] pp.6707–6723. Available at: <https://www.cs.cmu.edu/~sherryw/assets/pubs/2021-polyjuice.pdf> [Accessed 22 Sep. 2022]

Sundararajan, M., Taly, A. and Yan, Q. (2017). Axiomatic Attribution for Deep Networks. arXiv:1703.01365 [cs]. [online] Available at: <https://arxiv.org/abs/1703.01365>

Shrikumar, A., Greenside, P. and Kundaje, A. (2019). Learning Important Features Through Propagating Activation Differences. [online] Available at: <https://arxiv.org/pdf/1704.02685.pdf>

Ancona, M., Ceolini, E., Öztireli, C. and Gross, M. (2022). Towards better understanding of gradient-based attribution methods for Deep Neural Networks. [online] openreview.net.

Available at: <https://openreview.net/forum?id=Sy21R9JAW> [Accessed 22 Sep. 2022]

Lal, G.R. (2021). Explainable Neural Networks: Recent Advancements, Part 3. [online] Medium. Available at: <https://towardsdatascience.com/explainable-neural-networks-recent-advancements-part-3-6a838d15f2fb> [Accessed 22 Sep. 2022]