Tutorial objective

The goal of this tutorial is to build, validate, and explain a credit risk classification model using real-world tabular data. Students are expected to follow best practices in model development, demonstrate the added value of complex models over interpretable baselines, and apply post-hoc explainability methods to analyze model behavior at both global and local levels.

1. Perform an exploratory data analysis to assess data quality, feature distributions, missing values, correlations, and potential sources of bias relevant to credit risk modeling.
2. Define the prediction target and build a complete feature matrix suitable for both linear and tree-based models.
3. Train an intrinsically interpretable baseline model and justify its role as a reference (not necessarily the final model).
4. Train at least one high-capacity model and demonstrate a statistically meaningful performance gain over the baseline.
5. Validate the robustness of the selected model using cross-validation and overfitting diagnostics.
6. Freeze the final model and dataset split to prevent explanation leakage.
7. Compute global permutation feature importance and identify the top drivers of the model's predictions.
8. Compare global feature importance with the baseline model's intrinsic explanations and highlight discrepancies.
9. Use PDP to characterize the marginal effect of at least two numerical features and assess whether linear assumptions hold.
10. Use ICE to detect heterogeneous behaviors that are hidden by global averages.
11. Use the H-statistic to quantify interaction strength between `Credit amount` and `Duration`, and verify whether the model relies on non-additive effects.
12. Apply LOFO importance and explain cases where retraining-based importance disagrees with permutation results.
13. Train an interpretable surrogate model to approximate the black-box predictions and evaluate its fidelity.
14. Select a correctly classified instance and generate a local explanation using LIME.
15. Select a misclassified or borderline instance and analyze its LIME explanation.
16. Produce a final synthesis stating what aspects of the model behavior are reliably explainable and which remain opaque.