Final Project – XAI Case Study

Overview

The goal of this final project is to design and implement an **end-to-end Explainable AI (XAI) case study** on a problem and dataset of your choice. You will build or reuse a model, apply suitable explainability techniques, and use the resulting explanations to *do something useful*: improve the model, support a decision, or derive concrete domain insights.

The project is meant to be realistic and portfolio-worthy: at the end, you should have a small case study that you would feel comfortable presenting in a job interview.

Learning Objectives

By completing this project, you should be able to:

- Formulate a real-world ML problem where explainability is valuable.
- Select and train appropriate models (tabular, vision, time series, or text/LLMs).
- Apply XAI techniques you have learned in the course:
  - Tabular XAI: feature importance, SHAP, LIME, counterfactuals, . . .
  - Computer vision XAI: Grad-CAM, IG, occlusion, . . .
  - Time-series XAI: feature-based SHAP, temporal attribution, attention heatmaps, . . .
  - LLM/Text XAI: token-level saliency, example-based explanations, rationales, . . .
- Evaluate and criticise the explanations (sanity checks, faithfulness, stability).
- Turn explanations into **actionable recommendations** for model or domain stakeholders.

1. Problem and Dataset

You may choose any domain and data modality (tabular, images, time series, text, or multimodal), provided that:

- The task is clearly defined (input, output, objective).
- The dataset is publicly available or can be anonymised for academic use.
- There is a plausible real-world stakeholder who would care about both the model *and* its explanations.

In your report you must clearly describe:

- The **task** (e.g. credit scoring, defect detection, demand forecasting, activity recognition, toxicity detection, . . .).
- The **dataset**: origin, link, size, data types, and key preprocessing steps.
- The **stakeholder** and why explainability matters in this scenario.

2. Models and Evaluation

You must implement at least one main model to obtain explanations from that is **not overfitted**.

Your evaluation should:

- Use appropriate train/validation/test splits (time-aware for time series).
- Use relevant metrics (accuracy, F1, ROC-AUC, MAE/RMSE, BLEU, etc.).

3. Explainability Techniques

3.1. Option A

You must apply **at least three complementary XAI techniques**, covering:

- At least one **global** explanation (average model behaviour).
- At least one **local** explanation (specific predictions / instances).
- At least one **evaluation or sanity check** for explanations.

For each technique, justify:

- Why it is appropriate for your model and data.
- How you interpret the results (with plots and concrete examples).

3.2 Option B

You must develop **a new XAI technique with grounded theory** that improves explanations from the state of the art.

For the developed technique, justify how your technique is better compared to existing techniques using quantitative and qualitative experiments (explanation quality, computational time, memory management...).

4. Actionable Use of Explanations

Explanations in this project must be **actionable**: they should help you change or recommend something tangible. At least one of the following must be clearly demonstrated:

4.1 Model Improvement

Use explanations to:

- Detect data leakage or spurious correlations.
- Identify redundant or harmful features and remove them.
- Design better features (e.g. new lags, interactions, calendar variables).
- Adjust model architecture or regularization based on observed behaviour.

Show **before/after** metrics to demonstrate the impact.

4.2 Domain / Business Insights

Use explanations to answer domain-relevant questions, such as:

- What drives customer churn, risk, demand, or anomalies?
- How behaviour differs across groups, periods, or operating conditions?
- Which interventions (pricing, promotions, maintenance, prioritisation) are suggested by the explanations?

4.3 Fairness, Bias, and Robustness

If the data permit, you may focus on:

- Analysing differences in explanations across subpopulations.
- Revealing unwanted bias or instability.
- Proposing mitigation strategies and evaluating their effect.

5. Evaluation and Critique of XAI

You must critically reflect on the **quality and limitations** of your explanations. Possible elements:

- Compare different explanation methods: where do they agree or disagree?
- Perform sanity checks:
  - Randomise labels or model parameters and check whether explanations collapse.
  - Remove or shuffle features deemed “important” and observe performance changes.
- Discuss:
  - Where explanations might be misleading or unstable.
  - How a stakeholder might misinterpret them if not used carefully.

6. Deliverables

Your submission must include:

6.1 Code Repository

- A Git repository (e.g. GitHub, GitLab) with:
  - Clear structure (e.g. `notebooks/`, `src/`, `data/` or data download script).
  - A `requirements.txt` or environment file.
  - Instructions in a `README.md` on how to run the main experiments.

6.2 Notebooks / Scripts

- At least one Jupyter notebook / Python script showing:
  - Data loading and preprocessing.
  - Model training and evaluation.
  - XAI computations and visualisations.
  - Comments explaining key results.

This deliverable might change based on the MLOps project requirements.

6.3 Written Report (Max. 8 pages, 5–6 recommended)

The report should include:

- **Introduction**: problem description, motivation, stakeholders.
- **Data and Methods**: dataset, models, XAI techniques.
- **Results**: model performance, key explanations, visualisations.
- **Actions and Insights**: model improvements and/or domain recommendations.
- **Discussion**: limitations, risks, and ideas for future work.

7. Assessment Criteria

Here is the grading rubric (out of 10 points):

- **Problem choice and clarity (1.5 points)**: Is the task realistic, well-motivated, and clearly formulated?
- **Modelling and evaluation (1 point)**: Are baselines appropriate? Are evaluation protocols correct and justified?

- **XAI implementation (2.5 points)**: Are techniques appropriate, correctly used, and well visualised?
- **Actionable use of explanations (2 points)**: Are explanations used to improve the model and/or provide concrete domain recommendations?
- **Critical reflection and evaluation of XAI (2.5 points)**: Are limitations and potential pitfalls discussed thoughtfully?
- **Reproducibility and communication (0.5 points)**: Is the project easy to run and clearly explained?

8. Academic Integrity

You are encouraged to:

- Use open-source libraries and pretrained models.
- Be inspired by existing examples and tutorials.
- Use generative AI tools to aid in the project development.

However, you must:

- Clearly cite all external code and resources.
- Write your own analysis, explanations, and report.
- Understand all the steps you make during the project development.

Further questions might be done to the team's members to assess your understanding of the project (and grade might be affected).

The objective is not to build the most complex model, but to show that you can use Explainable AI thoughtfully to support better models and better decisions.