Project name: GTM (Go to market) Machine Learning Explainability.

Company: Schneider Electric S.A

Mentors: Mariana Serrão, Adriana Álvaro

SE Context:

Schneider Electric is a company in the Energy Management and Industrial Automation, energy transition and Industry 4.0 sector. It employs more than 130,000 people around the world and its markets are divided between Data Centers, Buildings, Industry and Infrastructure. In Iberia's Data team we launch initiatives and carry out projects throughout the Data ecosystem. From Process Automation, where through advanced analytics we improve existing processes with automations. Data architectures, where we ensure that the scalability of information storage is sustainable and well structured. Change management, where through proposals and initiatives we try to change the mentality and culture of the company towards Data-Driven. To advanced analytics, where through innovation projects and complex analysis, we use Machine Learning and Deep Learning techniques to seek insights and define the business strategy.

Description:

As a company that sells complex solutions and projects, we generate millions of data points about past sales **opportunities** that are stored in our CRM. In the Strategic Advanced data analytics team, our mission is to make better decisions based on patterns hidden in the data.

🎯 Datathon Challenge: Explainability on Classifier Models

📌 Objective

Your mission is to **train a binary classification model** to predict whether an opportunity (a sale) is going to be won or lost. Then apply **Explainability techniques** to understand how each feature influences the predictions.
✅ The dataset is already prepared — no need for extra transformations or pipelines, is ready to train.

Focus (in terms of time):

- **15%** → Train the model
- **85%** → Explain the predictions

Once the model reaches a reasonable performance and predicts whether an opportunity is **won** or **lost**, the real challenge begins: **Why did the model make that decision?**

📁 Dataset Overview

Historical data related to opportunities

The dataset includes the following columns:

- **id**
  Unique identifier for each observation. Not predictive — do not use as a feature.
- 🎯 **target_variable**
  Binary target variable:
  - 1 → Opportunity won
  - 0 → Opportunity lost
- 📊 **Other Features**
  At the end of the document you will find a list of the features and their explaination

---

🛠 Challenge Steps

1. **Train an classification model**
   - Build a binary classification model using the dataset.
   - Achieve a reasonable performance. The better the performance, better the possible explanations. Min performance 0,7 f1 score.
2. **Apply Explainability Techniques**
   - Examples of techniques (you don't need to use all of them, they are just ideas):

   SHAP/Shapley values,

   Local Interpretable Model-agnostic Explanations (LIME)

   Partial Dependence Plots (PDP)

   Accumulated Local Effect (ALE)

   Model built-in feature importance.

   - The objective is to find:
     - 🌍 **Global insights**: Which features matter most overall?
     - 🔍 **Local insights**: Why did the model predict a specific case as won or lost?

ℹ️ **Note:**
This dataset is **simplified** (around 15 variables), making interpretation easier.
However, in real-world scenarios with **100+ features**, interpreting SHAP/other explainability plots can become challenging.
💡 **Idea:** Use a **Large Language Model (LLM)** to help summarize and interpret SHAP/other explainability outputs automatically, turning complex patterns into human-readable insights.

3. **Deliverables**
   - A short report or presentation including:

        - ✅ Model performance summary
        - 📈 Explainability technique proposal and how to exploit it as a user
        - 💡 Insights: Why does the model predict an opportunity as won or lost?

🏆 Evaluation Criteria, weight of each part on the overall evaluation

- Model performance: 25%
- Use of appropriate explainability techniques: 30%
- User friendly insights. This means a non technical person should be able to understand the results: 30%
- Creativity: 15%
- EXTRA 5%: what would have you done if you had had more time to implement?

Resources:

Explainability

https://christophm.github.io/interpretable-ml-book/

Data dictionary:

- **product_A_sold_in_the_past**: Historical sales of Product A with the customer concerned.
- **product_B_sold_in_the_past**: Historical sales of Product B with the customer concerned.
- **Product_A_recommended**: Indicates whether Product A was recommended in the past to the customer concerned.
- **product_A**: amount we are trying to sell of product A in this opportunity
- **product_C**: amount we are trying to sell of product C in this opportunity
  product_D: amount we are trying to sell of product D in this opportunity
  cust_hitrate: Customer success rate in previous interactions.
- **cust_interactions**: Number of interactions with the customer.
- **cust_contracts**: Number of contracts signed with the customer.
- **opp_month**: Month when the opportunity was created.
- **opp_old**: indicates if the opportunity has been open for a lot of time.
- **competitor_Z**: Presence of competitor Z in the opportunity.
- **competitor_X**: Presence of competitor X in the opportunity.
- **competitor_Y**: Presence of competitor Y in the opportunity.