SEB DATA CHALLENGE

Welcome to the SEB Data Challenge! Your goal is to solve a small task that is similar to the kind of problems we deal with at SEB and show us a glimpse of your technical skills.

Guidelines:

- The task consists of two parts with small questions. However, you are not required to answer all of them or to provide a perfect solution. Feel free to scope the assignment as you consider appropriate.
- **Note that the focus is not on accuracy.** We are interested in seeing how you reason, handle the data, your technical soundness and coding skills.
- We recommend using Python though you are allowed to use other programming languages. Bonus points for using tools like Spark, Docker or Github.
- The submission must include:

    - The code to reproduce the results (script, notebooks/markdown, etc.). You will present this to us as a technical audience.
    - A presentation with a summary of the setup, the steps taken and the results. **You must show this to us as if we are non-technical business stakeholders and you are presenting results to promote your model.**

Data:

You can find the dataset in a compressed file in the following link:

https://drive.google.com/file/d/18bCjjmWvHpsP4r9w2_RB1kJmFVZnssEv/view?usp=sharing

The dataset consists of:

- ***customers.csv*** file with columns:

CLIENT_ID	Customer identifier
ACCOUNT_ID	Account identifier
GENDER	Customer gender
BIRTH_DT	Birth date (YYYYMMDD)

ACTIVE	Active customer flag (1=active, 0=inactive)
LOAN	Flag indicating if the customer was granted a loan (1=yes, 0=no)
DISTRICT_ID	District identifier
SET_SPLIT	Dataset split (train or test)

- ***transactions.csv*** file with columns:

TRANS_ID	Transaction identifier
ACCOUNT_ID	Account identifier
DATE	Transaction date (DDMMYYY)
AMOUNT	Transaction amount
BALANCE	Account balance
TYPE	Transaction direction
OPERATION	Type of operation involved

- ***districts.csv*** file with columns:

DISTRICT_ID	District identifier
N_INHAB	No. of inhabitants
N_CITIES	No. of cities
URBAN_RATIO	Ratio of urban inhabitants
AVG_SALARY	Average salary
UNEMP_95	Unemployment rate 1995
UNEMP_96	Unemployment rate 1996
N_ENTR	No. of entrepreneurs per 1000 inhabitants
CRIME_95	No. of committed crimes 1995
CRIME_96	No. of committed crimes 1996

Questions:

A. Data Exploration:

The first task is to explore the data and extract insights about the customers. Visualise your findings in your presentation to business stakeholders. This step is the most important for us to understand your reasoning and business communication skills and how you structure a project.

B. Predictive model:

Build a model to predict which customers were granted a loan (binary classification). Use the column LOAN as the target and the column SET_SPLIT to break down the data into train and test sets.

- What are the most important features in the model?
- How does the model performance compare in the train and test sets?
- What would you do to improve the model if you had more time?