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 model accuracy. We are interested in seeing how you reason, 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.)
    - A **presentation** with a summary of the setup, the steps taken and the results. Maximum 5 slides. Include references if applicable.

Data:

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

https://goo.gl/8rJPMx

The **dataset** consists of:

- *Customer.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 (YYYYMMDD)
AMOUNT	Transaction amount
BALANCE	Account balance
TYPE	Transaction direction
OPERATION	Type of operation involved

- *District.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 entrepeneurs per 1000 inhabitants
CRIME_95	No. of commited crimes 1995
CRIME_96	No. of commited crimes 1996

Questions:

(A) Data exploration:

The first task is to explore the data and extract insights about the customers.

Potential questions to consider:

- How many transactions did an average customer complete in the period? How much did they spend? Does it change over time?

- Do different customer profiles show different behavior? Is the transaction pattern homogeneous across geographic regions?
- Visualize one of your findings

(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?

(C) A/B testing (optional bonus question)

Assume the bank is to launch a campaign to offer loans to their customers and you are assigned with finding the optimal channel between two options: email or mobile app notification.

To do so, you decide to run a experiment with a single factor and two levels (A/B test) in the initial part of the campaign.

- How would you select which customers are in the email group or in the mobile app group in the test?
- How would you conclude which channel is optimal at the end of the test? What is the minimum sample size?