Data Science Digital Race 2025
Round 2 Question

Round 2: Applied Machine Learning Task

Duration: 4 hours
Maximum Points: 60
Platform: Google Colab

DATASET INFORMATION

You are given a dataset about bank customer churn. Your objective is to build models that predict whether a customer will churn (leave the bank) or not.

- Dataset size: ~10,000 customers
- Target variable: churn
    - 0 = Customer stays
    - 1 = Customer churns

- Features:
    - customer_id – Account Number
    - credit_score – Credit Score
    - country – Country of Residence
    - gender – Categorical (Male/Female)
    - age
    - tenure – Years as a bank customer
    - balance – Account balance
    - products_number – Number of bank products
    - credit_card – 1 if customer has a credit card, else 0
    - active_member – 1 if active, else 0
    - estimated_salary – Yearly salary

INSTRUCTIONS

1. You will work with a dataset on **bank customer churn**.

2. Your notebook must include:
    - **Code cells** for data loading, preprocessing, model training, evaluation, and visualization.
    - **Markdown cells** for written answers to the guided questions, explanations of your approach, and interpretation of results.

3. Clearly label each answer in Markdown according to the task number (e.g., *Answer A1, Answer B3*).
   Your notebook must:
    - Run without errors from start to finish.
    - Contain both code and written explanations.

4. Submission must be made **before the round ends**. Any submission after the time limit will not be accepted.

5. Save your notebook using the following file naming format:
   DSDR_R2_<TeamNameOrYourName>.ipynb (e.g., DSDR_R2_UMDAC.ipynb).

6. You may use up to two hints:
    - First hint available after 2 hours.
    - Second hint available after 3 hours.
    - Each hint used deducts 5 points from your maximum score.

7. The use of large language models (LLMs) or other AI tools to generate answers is not allowed and any submission found using them will be disqualified.

TASKS

Part A: Data Understanding (10 points)

1. What is the overall churn rate in the dataset?
2. Provide the distribution of customers by country.
3. Compare the average credit score between churned and non-churned customers.

Part B: Model Development (25 points)

4. Preprocess the dataset appropriately (e.g., categorical encoding, scaling).
5. Build at least **two different classification models** (e.g., Logistic Regression, Random Forest, XGBoost).
6. Evaluate each model using:
    - Accuracy
    - Precision
    - Recall
    - F1-score
7. Compare and discuss which model performs better overall.

Part C: Feature Insights (10 points)

8. Perform feature importance analysis using one of your models.
9. Create four distinct age group categories from the dataset and analyze how customer churn differs across these groups. In your answer, you should:
    - Clearly state the age ranges used for grouping.
    - Calculate the churn rate for each age group.
    - Identify the age group with the highest churn risk and provide a brief explanation.

Part D: Model Validation & Interpretation (15 points)

10. Suppose the bank is more concerned about identifying churned customers rather than keeping accuracy high. Which evaluation metric would be the most important in this context? Explain your reasoning.
11. Based on your models, what are two key factors that strongly influence customer churn?

Submission Requirements

- Submit your **Google Colab Notebook (.ipynb)** before the round ends.
- Ensure your notebook can run fully from top to bottom without errors.
- Clearly label answers to each question in Markdown.