DATA SCIENCE CODING CHALLENGE

Dear candidate, as part of the recruitment process, we invite you to complete a technical challenge to help us evaluate your skills:

PROBLEM STATEMENT

Nexus Insurance, a subsidiary of the **Nexus Group**, is a multinational non-life insurer with a large **motor vehicle insurance portfolio**. The company records detailed transaction-level data for each policy, including renewals, premium changes, vehicle characteristics, and claims information.

You have been asked to analyse this historical portfolio and explore how data and machine learning can support **pricing analysis, risk understanding, and portfolio decision-making**.

The dataset provided represents **policy transaction data** for a motor insurance portfolio over **three years**:

- Each row corresponds to a policy transaction or renewal
- The same policy may appear multiple times across years
- Premiums may change over time even in the absence of claims
- Claims are relatively rare, as expected in motor insurance

The dataset reflects **real operational behaviour**, not a simplified academic dataset.

Dataset to be downloaded from: Dataset of motor vehicle insurance portfolio

Task A – Policy-Level Pricing Assessment

Develop an analytical component that evaluates historical motor insurance policy data to **analyse and model premium behaviour over time**.

The solution should:

- Correctly handle the longitudinal nature of the data, including multiple records per policy
- Identify and analyse key patterns affecting premiums across policies and renewals
- Frame and implement a single, well-defined machine learning task related to pricing or risk
- Produce interpretable results that explain **what factors influence premium variation**

The output must be suitable for **portfolio analysis and pricing review**, with clear justification of model choices and limitations.

Task B – Investigative Decision Support (Optional)

Design a **lightweight decision-support component** that operates on an individual policy.

This component should:

- Consume one policy record and the output from Task A
- Apply a small set of transparent, data-driven rules
- Produce a concise, structured summary **for the pricing or risk of the policy**
- This component should **highlight features leading to predicted premium or risk.**
- Avoid speculative and unsupported reasoning (control hallucinations of the LLM)

The output must be **structured, explainable, and suitable for rapid review** by analysts.

EXPECTED DELIVERABLES

- The application must be written in **Python**. Code should be clear, well-structured, and readable.
- A **reusable pricing or risk analysis module** that evaluates motor insurance policies and produces model outputs suitable for portfolio or policy-level review.
- An optional **lightweight decision-support component** that generates structured outputs for a single policy using model results and simple rules.
- A brief **report or README** explaining assumptions, evaluation choices, and instructions to run the solution locally.
- A **link to a code repository** (e.g. GitHub) containing the proposed solution. The submission should reflect sound software engineering practices, including clean structure, clarity, and consideration for scalability of the overall machine learning solution.
- A **high-level description of the end-to-end machine learning system design**, covering the process from development through deployment and monitoring. This description should focus on architectural and conceptual aspects rather than code-level specifics and will be discussed during the interview alongside the submitted solution.