**Data Science Interview Task:**

Predicting Insurance Risk

# Problem Statement

The following theoretical problem is designed allow you to showcase your Data Science skills and industry knowledge.

We have been approached by the business to use machine learning to determine the claim risk associated with a policy. There are people that are dubious about the use of modern machine learning techniques over traditional risk quantification methods. We need to showcase how and why modern machine learning should be adopted, and the modelling approach and outputs should be easily explainable to stakeholders.

The goal of this task is to predict the risk of offering an insurance policy to a potential customer, based on their personal and financial information. The risk is defined as the probability of the customer filing a claim within a year of purchasing the policy. A higher risk means a higher chance of the customer making a claim, which would result in a loss for the insurance company. Therefore, the insurance company wants to understand the risk profile of each customer to determine the level of risk currently held, and how to price products. The business would also like to understand how we could use this model to ensure that only 5% of the people they sell a policy to would claim on their policy.

# Deliverables

As a proof-of-concept to address this problem, we would like you to produce a self-contained Jupyter notebook which demonstrates your understanding of and approach to the problem, along with a potential solution. Please include any other files required for interacting with what you produce.

In this theoretical scenario, your notebook will be shared with a broad audience, including stakeholders who may not be familiar with the full context of the task. While you can assume readers are comfortable with Python and Jupyter notebooks, you should not assume they have any prior knowledge of your specific approach.

Your notebook should be self-contained and well-annotated, clearly explaining:

* **What you did:** Describe the steps you took, including data cleaning, feature engineering, modelling, evaluation, etc.
* **Why you did it:** Justify your decisions—why you chose a particular method, metric, or approach over alternatives.
* **What you considered but didn’t do**: If you explored or ruled out other options, briefly explain your reasoning.
* **What the results mean:** Interpret your findings in a way that would be understandable to someone reviewing the notebook without additional context.

Think of this as an opportunity to demonstrate not just your technical skills, but also your ability to communicate your thought process and understanding of the problem.

#### Data Exploration and Preprocessing

We want to understand how you approached preparing the data for analysis or modelling. Your notebook should clearly demonstrate:

* **Exploration:** How did you initially explore the dataset? What patterns, anomalies, or issues did you identify?
* **Pre-processing Steps:** What transformations or cleaning steps did you apply?
* **Rationale:** For each major step, briefly explain why you chose that approach. What alternatives did you consider, and why did you rule them out?
* **Assumptions:** What assumptions are you making about the data or the problem? How might these affect your results?

#### Modelling

Once the dataset is analysed and prepared for modelling, we'd like you to utilise machine learning to build, evaluate and present a model which could be used to tackle the above problem. The data which has been provided is synthetic, and this task reflects a proof-of-concept development, so the technical performance won't be accounted for when reviewing the task. However, we still want to understand your process and technique for model training and development.

With insurance being a regulated industry, the business needs to be comfortable with what's been created - especially given the scepticism to Machine Learning. Given this, explainability is paramount!

# Questions

The following questions are to understand how you would account for the solution outside of the developed model. Please include written answers to these questions in your response, either within your notebook, or in a separate document (preferably in a markdown or word editor format, e.g. .md or .docx).

* Assuming the business likes your model, what are the next steps to go from the Jupyter Notebook you've created to a productionised service? Do you have experience with doing this previously?
* What considerations are the for ensuring that if you productionised this model, the business could leverage it?
* What are the steps you would take to provide this to the business? What assumptions are you making for this to be possible?
* Assuming that the team you're working in only consists of data scientists, which other teams in the business would you need to speak to?
* What is in scope and out of scope for your responsibility?