1. Briefing

This exercise is designed to assess the approach you would take to developing a machine learning model to answer a business question.

The scenario and assets have been simplified to shorten the process so it can fit into a few hours, where it would normally take a team several weeks to complete. With this in mind, we ask you to suspend some disbelief where you know some aspects would be more complex or work differently in real life.

Any code that is written in completing the exercise should be done in Python and in a presentable form. If it is not possible for you to use Python for this exercise please contact your interviewer as soon as possible.

Please prepare a 10-minute presentation covering the aspects of the exercise and submit the presentation along with your Python code and any assumptions you have made by the deadline communicated to you.

2. Scenario

You have been asked by the marketing department to help them identify existing customers who would be likely to take out a home insurance product.

A previous campaign was conducted 6 months ago, where the marketing team targeted a random selection of your mortgage customer base. These customers were offered this home insurance product but only a small number of them took it out. The campaign results are available to you, and you also have access to other data you already hold about customers.

3. Goal

The company wants to increase the number of customers who take out an additional home insurance product.
The marketing department has asked you to use the previous campaign data to improve the selection of candidate customers so a higher proportion will purchase the product in a future marketing campaign.

4. Questions to think about

What data is available? (Data discovery and sourcing)
What data sources do we have?
What information do those sources contain that might be useful?
Are we able to use the data for this purpose?

How good is the data? (Data quality)
Are there missing values, different units, repeated columns?
How can we join the different datasets we need?

Are there any obvious patterns in the data? (Exploratory analysis)

How could we build a model? (Model and feature selection)
What types of model could we try?
What features do we want to use?

How do we evaluate the output of the model?
What measures should we look at?
What thresholds are we targeting?

How can we explain the model?
If the marketing department, a customer or the regulator want to understand why we designed the campaign, what rationale can we provide?

5. Data Available

Campaign Dataset

- Participant ID
- Title
- First Name
- Last Name
- Age
- Postcode
- Marital status
- Education qualification level
- Number of years in education
- Job title
- Familiarity with Nationwide (1 = don't know it, 10 = a big fan, follow on social media)
- View of Nationwide brand (1 = really dislike, 10 = really like)
- Interested in insurance product (0=No, 1=Yes)
- Email address
- Did they buy the product?

Mortgage Dataset

- Full name (first, middle and last name combined)
- Date of birth
- Town of residence
- Employer PAYE reference
- Salary (can be a yearly salary, monthly salary, weekly salary, range of a yearly salary, yearly salary in a foreign currency)
- Time with current employer (years, integer)
- Time with current employer (months, integer < 12)
- Length of working week (hours)
- Capital gain
- Capital loss
- New mortgage (mortgage started in the last 3 months)
- Sex
- Religion
- Relationship
- Race
- Native country
- Work class
- Demographic characteristic (a numerical value where close values mean similar demographics)