## **Project details - classification**

**Background:**

You are working as a risk analyst with a bank. Apart from the other banking and loan services, the bank also provides credit card services, a very important source of revenue for the bank. The bank wants to understand its customers’ demographics and other characteristics that accept a credit card offer and that do not accept a credit card.

Usually, the observational data for these kinds of problems is somewhat limited in that often the company sees only those who respond to an offer. To get around this, the bank designs a focused marketing study, with 18,000 current bank customers. This focused approach allows the bank to know who does and does not respond to the offer, and to use existing demographic data that is already available on each customer.

**Objective:**

The task is to build a model that will provide insight into why some bank customers accept credit card offers. There are also other potential areas of opportunities that the bank wants to understand from the data.

Your senior management has also posted these other questions that will help them better understand their customers.

**Data:** The data set consists of information on 18,000 current bank customers in the study. These are the definitions of data points provided:

* **Customer Number**: This is a sequential number assigned to the customers (this column is hidden and excluded – this unique identifier will not be used directly).
* **Offer Accepted**: Did the customer accept (Yes) or reject (No) the offer. Reward: The type of reward program offered for the card.
* **Mailer Type:** Letter or postcard.
* **Income Level**: Low, Medium, or High.
* **Bank Accounts Open**: How many non-credit-card accounts are held by the customer.
* **Overdraft Protection**: Does the customer have overdraft protection on their checking account(s) (Yes or No).
* **Credit Rating**: Low, Medium, or High.
* **Credit Cards Held**: The number of credit cards held at the bank.
* **Homes Owned:** The number of homes owned by the customer.
* **Household Size**: Number of individuals in the family.
* **Own Your Home:** Does the customer own their home? (Yes or No).
* **Average Balance:** Average account balance (across all accounts over time). Q1, Q2, Q3, and Q4
* **Balance:** Average balance for each quarter in the last year

### **Exploring the data**

We encourage you to thoroughly understand your data and take the necessary steps to prepare your data for modeling before building exploratory or predictive models. Since this is a classification model, you must use logistic regression for classification for building a model.

To explore the data, you can use the techniques that have been discussed in class. Some of them include using the describe method, checking null values, using \_matplotlib\_ and \_seaborn\_ for developing visualizations.

The data has a number of categorical and numerical variables. Explore the nature of data for these variables before you start with the data cleaning process and then data pre-processing (scaling numerical variables and encoding categorical variables).

For the target variable (Offer accepted – Yes/No), it is also important to check the data imbalance ie the number of people who responded with a yes vs the number of people who responded with a no.

You will deep dive into the data for customers who accepted the offer vs the customers who did not and check their characteristics. For e.g., we select the **Yes** level in **Offer Accepted** and then examine the distribution of accepted offers across the other variables in our data set and similarly for people who did not accept the offer.

### **Model**

Compare the error metrics in `train` and `test` sets to make sure that your model doesn't suffer from overffiting/underfitting and find the scaler that best fits your data ( you are free to try other scalers not covered in the course ).