**Problem Statement**

By now, you understand the process of building a response model on the bank marketing dataset. This assignment is similar, with a slight twist.

 To reiterate the response model problem discussed in the lectures — you wanted to predict the probability of response of each prospect and target the ones most likely to respond to the next telemarketing campaign. The steps were as follows:

1. Identify the relevant predictor variables for response using EDA
2. Build predictive models and choose the best one
3. Sort the prospects in decreasing probability of response (predicted by the best model) and target the top X% (or top Y deciles), where X would be determined by your business objective (e.g. maximising the overall response rate/the number of responders at a fixed marketing cost)

 When you look at the important variables included in the final model, you will see a variable ‘**duration**’ — the duration of the phone call in seconds. When you present this model to the Chief Marketing Officer (CMO), she would note that ‘duration’ has a positive correlation with the response, which is a problem for the marketing team. As the duration increases, the cost of telemarketing increases linearly, and that is the last thing they want.

 There are two problems with having the variable ‘duration’ in the model:

* When the marketing team procures prospect data, ‘duration’ is not present in it, since the call hasn’t been made yet
* In your analysis of marketing cost and response, you had assumed that the cost of a phone call is independent of duration (₹1 per call) — which is not true

## Tasks

To solve these problems, you and the CMO should decide to build another model without the variable ‘duration’. That will help you understand the relationship of other variables with the response.

You also decide that the business objective is to achieve 80% of total responders at the minimum possible cost. The total number of responders is the total number of prospects who responded, from the available data of about 45,000 data points.

To find the number of prospects you should target (i.e. how many top deciles you should target), you will assume that the cost of a call varies with duration as follows:

**Cost per call (INR)** = 0.033\*(duration\_in\_seconds) + 0.8

 Based on this, you will figure out the X in top X%, i.e. how many prospects should be called to meet the business objective.

The checkpoints are given below:

## Checkpoints

**Note**: Before starting, you will find it helpful to create a unique ID for each prospect. Also, you only have to submit an R file for this assignment, so you have to **report some of the metrics as comments** in the R file.

1. Data preparation (no marks assigned for this step)
   * You can use the code provided in the lectures to complete all data preparation steps
2. Build a logistic regression model without using the variable 'duration'
   * Perform variable selection using the usual methods
   * Sort the data points in decreasing order of probability of response
   * Find the optimal probability cut-off and report the relevant evaluation metrics
3. Create a data frame with the variables prospect ID, actual response, predicted response, predicted probability of response, duration of call in seconds, and cost of call
   * While creating the data frame, calculate the cost of call for each prospect in a new column
4. Find the number of top X% prospects you should target to meet the business objective
   * Report the average call duration for targeting the top X% prospects to the CMO (**report this as a comment** in the R file)
5. Create a lift chart
   * The x-axis contains the number of prospects contacted; the y-axis contains the ratio: response rate using the model/ response rate without using the model