Problem Statement

Business Understanding - Bank Marketing:

A Portuguese bank had conducted a telemarketing campaign for a term deposit product somewhere around late 2010. A term deposit is very similar to a fixed deposit, where you deposit money for a fixed period of time.

Through the campaign, the bank had collected data about the prospects' demographics, other financial products they have purchased in the past (loans, deposits, etc.), the number of times they were called, etc. They also recorded the response data, i.e., whether the person had subscribed to the term deposit product, which is the target variable.

The bank's marketing team wants to launch yet another telemarketing campaign for the same product. You, an analyst at the bank, want to answer the following questions using the past data:

1. Which prospects are more likely to buy the product (i.e., to respond)?
2. Which attributes determine the propensity to buy a term deposit?
3. Once you predict the likelihood of response, how many prospects should you target for telemarketing?
4. By how much can you reduce the marketing cost using the model, and how many prospects will you acquire?

Let's see how you can answer these questions using what are called 'response models' in marketing analytics.

<u>Problem statement:</u>

We want to predict the probability of a response from each prospect and target the ones most likely to respond to the next telemarketing campaign. The steps were as follows:

1. Identify relevant predictor variables for a response using EDA.
2. Build predictive models and choose the best one.
3. Sort the prospects in order of 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/number of responders at a fixed marketing cost).

In the dataset we have a variable called ‘Duration’. Now there are problems with including the variable ‘duration’ in the model:

- The prospect data procured by the marketing team does not contain ‘duration’, since the call has not been made yet.

Tasks

To solve these problems, you should resolve to build model without including the variable 'duration'. This will help you understand the relationship of the other variables with the response.

Also, set the business objective to achieving 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 prospects.

Based on this information, calculate the X in the top X%, i.e., how many prospects should be called to meet the business objective. Further, create a presentation for the CMO highlighting the results and the methodology employed.

Checkpoints

The checkpoints for the assignment are as follows:

1. Perform data preparation.
2. Build a logistic regression model without using the variable 'duration'
    - Select variables using the usual methods
    - Sort the data points in decreasing order of the 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 the call in seconds and cost of the call
    - While creating the data frame, list 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
5. Create a lift chart
    - The x-axis should show the number of prospects contacted; the y-axis should show the ratio of the response rate using the model and the response rate without using the model
6. Determine the cost of acquisition
    - Consider cost = 1*number of contacts made in the current campaign; determine the cost incurred for acquiring 80% of customers using the predictive model
7. Create a small presentation for the CMO highlighting your findings and the methodology used.

Data

Customer data: Demographic data, data about other financial products like home loans, personal loans, etc.
Campaign data: Data about previous campaigns (number of previous calls, number of days since the last call was made, etc.)
Macroeconomic data
Target variable: Response (Yes/No)