We have chosen the Bank Marketing Data Set from UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#>). The data is originally sourced from a Portuguese bank performing direct marketing campaigns. Upon these phone calls, a number of attributes from potential clients were captured.

Age – Age is a numeric variable that ranges from ages 18-95 years.

Job – Job type is a categorical variable that entered the client into types of occupations. The choices were: admin., unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician and services.

Marital – Marital status is a categorical variable that describes if the client is married, divorced or single. For widowed clients, the bank has selected the divorced selection.

Education – Education represents the client’s highest level of education. The categorical variable has the following selections: unknown, primary, secondary or tertiary. Primary level is equivalent to the US elementary and middle school grades. Secondary is similar to the US high school, but is split between more professional teachings versus vocations. Tertiary level represents all level of higher education from special licenses, bachelors to post doctorate school (<https://www.scholaro.com/pro/Countries/Portugal/Education-System>).

Default – Default represents if the client has defaulted in their credit history as a binary variable denoted with yes for default or no for no default.

Balance – Balance is a numeric variable

Housing – Housing is a binary variable which represents if the client has a home loan or not. The selections are yes or no.

Loan – Loan is similar to the housing variable except it represents if the client has a personal loan.

Contact – Contact is a categorical variable describing the method of telecommunication between the bank and the client. The selections are: unknown, telephone or cellular.

Day – Day is a numeric variable that represents the day of the month of the last contact.

Month – Month is represented as a categorical variable where each month is denoted by the first three letters. Every month of the year is represented. The month corresponds to the last contact made with the client.

Duration – Duration is the amount of time (a numeric variable) in seconds that the last contact with the client lasted. This attribute is not the total time the clients were interviewed. 1% of the instances lasted 10 seconds or less. Is there a reason why something like this is a cause of concern? Since they tell us there are multiple calls, perhaps it doesn’t matter.

Campaign – Campaign is a numeric variable which represents the number of times the bank has contacted, using any method of communication, the client during this particular marketing campaign.

pDays – pDays is a numeric variable which denotes the number of days that has passed between a previous marketing campaign and the current marketing campaign. The ranges are from -1 identifying these are new clients to the marketing campaign up to over 2 years ago.

Previous – Previous is a numeric variable that identifies the number of times a client has been involved with a marketing campaign with this bank prior to the current campaign. The ranges for this are 0 which would represent this is the first marketing campaign the client is involved with to 275. (all the data points are from 0-58 with 275 appearing to be an outlier?? Haven’t checked anything else here, just a notation at this point)

pOutcome – pOutcome represents the outcome of a previous marketing campaign. The categorical variable is identified by: unknown, other, failure or success. Each client that has not participated in a previous marketing campaign has been marked as unknown for pOutcome.

Y – Y signifies yes if the customer accepted a term deposit during the current marketing campaign. The binary variable is defined as yes or no.