**Data**

My proposed methods performance were assessed by using the real data from the UCI Machine Learning Repository (Moro et al., 2014). The dataset was obtained from a Portuguese banking institution from May 2008 to November 2010, and the marketing campaigns were based on phone calls. It was often the case that more than one contact to the same client was required to obtain whether the bank term deposit would be (‘yes’) or not (‘no’) subscribed. The missing values in this dataset are named as “unknown”, “non-existent”, and “999” in different variables. The dataset involves 41,188 phone contacts in total with 20 input variables and 1 output variable, which will be listed in Table 1. There are two types of input variables, which are numerical and categorical, and details are listed below. The classification goal is to predict if the customer will or not (yes=1/no=0) subscribe to the term deposit (response variable). The final data I used after cleaning up for missing values and dropping “duration” variable remaining 1,310 observations and 20 variables. The removal of “duration” could be traced from the data contributors’ descriptions of that variable. They noted “this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model (Moro et al., 2014).”

**Table 1. Variable Descriptions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Number** | **Variable Name** | **Description** | **Type** |
| 1 | Age | Age of the customer | numeric |
| 2 | Job | Type of job ('admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown') | categorical |
| 3 | Marital | Marital status ('divorced','married','single','unknown'; note: 'divorced' means divorced or widowed) | categorical |
| 4 | Education | Education status('basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown') | categorical |
| 5 | Default | Has credit in default? ('no','yes','unknown') | categorical |
| 6 | Housing | Has housing loan? ('no','yes','unknown') | categorical |
| 7 | Loan | Has personal loan? ('no','yes','unknown') | categorical |
| 8 | Contact | Contact communication type ('cellular','telephone') | categorical |
| 9 | Month | Last contact month of year ('jan', 'feb', 'mar', ..., 'nov', 'dec') | categorical |
| 10 | Day\_of\_week | Last contact day of the week (categorical: 'mon','tue','wed','thu','fri') | categorical |
| 11 | Duration | Last contact duration, in seconds. Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model. | numeric |
| 12 | Campaign | Number of contacts performed during this campaign and for this client (includes last contact) | numeric |
| 13 | pdays | Number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted) | numeric |
| 14 | Previous | Number of contacts performed before this campaign and for this client | numeric |
| 15 | poutcome | Outcome of the previous marketing campaign ('failure','nonexistent','success') | categorical |
| 16 | emp.var.rate | Employment variation rate - quarterly indicator | numeric |
| 17 | cons.price.idx | Consumer price index - monthly indicator | numeric |
| 18 | cons.conf.idx | Consumer confidence index - monthly indicator | numeric |
| 19 | euribor3m | euribor 3 month rate - daily indicator | numeric |
| 20 | nr.employed | Number of employees - quarterly indicator | numeric |
| 21 | y | Has the client subscribed a term deposit? | (binary: 'yes','no') |

**References**

Moro, Sérgio & Cortez, Paulo & Rita, Paulo. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems. 62. 10.1016/j.dss.2014.03.001.

Morgan J. N., Sonquist J. A. (1963) Problems in the analysis of survey data, and a proposal, Journal of the American Statistical Association, Vol. 58, Issue 302, 415-434.

General comments:

I think you are making nice efforts in obtaining data that could potential workable. However, the data description feels “uncomplete” and lack of some logical progress. You may want to state your research questions first, and then explain how the data will help answer the questions. You should also report the summary statistics of your choice of variables. It seems that you have done some data cleaning. Any suspicious issues? What is the total number of observation before you clean? 41000? Dropping down to 1310 seems rather dramatic. Would you worry about representativeness? We need to learn more…