Part 1. Homework 2: Logistic Regression

Write a report that contains the results of the computations that you are asked to carry out below, as well as the explanation of what you are doing. The main text should include pieces of source code and graphical and numerical output. Upload your answers in a .pdf document (use LaTeX or R Markdown, for instance), as well as the source code (*.R or *.Rmd, for instance). Your work must be reproducible.

1. Generalized Linear Model for JYB data

The Banking Entity Just-Your-Bank (JYB) has the purpose of reducing the telemarketing costs by decreasing the number of calls to clients not likely to buy the product.

The JYB dataset contains information of 28.645 calls. This is the list of the available variables:

- Client attributes

```
id: Customer ID
age: age in years
job:
    (admin., blue-collar, entrepreneur, housemaid, management,
    retired, self-employed, services, student, technician,
    unemployed, unknown)
marital: Marital status
    (Divorced, married, single, unknown)
education: Education level
    (basic.4y, basic.6y, basic.9y, high.school, illiterate,
    professional.course, university.degree, unknown)
default: is he/she a defaulter?
    (No, yes, unknown)
housing: does he/she has a mortgage?
    (No, yes, unknown)
loan: does he/she has a personal loan?
    (No, yes, unknown)
```

- Call attributes

```
contact: phone type (cellular, telephone)
month: month of the call
day_of_week: day of the call
    (mon, tue, wed, thu, fri)
```

- Campaign attributes

```
campaign: Number of contacts made this campaign for this client
    (including the current one)
pdays: number of days that have passed since the customer was contacted
    for the last time for a previous campaign
    (999 means that it was not previously contacted)
previous: number of calls made to this client before this campaign
poutcome: previous campaign result
    (failure, nonexistent, success)
```

- Indicators of the social and economic context

```
emp.var.rate: employment variation rate (quarterly)
cons.price.idx: Consumer Price Index (monthly)
cons.conf.idx: Consumer confidence index (monthly)
euribor3m: euribor a 3 mesos (daily)
nr.employed: number of employed (quarterly)
```

- Response variable

```
Y: The customer subscribed the deposit? (yes,no)
```

1. Do the Exploratory Data Analysis for this dataset. Make some data cleansing and missing data imputation (if adequate).Explain the most interesting conclusions.
2. With the original variables, fit the complete model without interactions and using the logit link function.
3. Evaluate possible first order interactions (between two factors or between a factor and a covariable) and include them in the model (if there were any).
4. Perform an automatic variables selection basen on the AIC & BIC. Make a comparison of the models and argue which one is chosen.
5. Validate the model by checking the assumptions
6. Interpret the final model