Predictive Modelling, Application to Bank Telemarketing

This case is based around a real-world dataset about telemarketing calls made by a Portuguese bank. You can find more information about this dataset here:

https://archive.ics.uci.edu/ml/datasets/bank+marketing

The bank is interested in a predictive model because it will allow them to call the right customers at the right times. From an analytics perspective, the primary distinguishing feature in this case is that solving a predictive problem is directly useful to the firm.

Please hand in both documents and outputs preferably using **Rmarkdown**.

Basic Explanatory Analysis

1. Load the data contained in the file data_telebank.csv and name the variable `dta_bank`
2. In one sentence, describe variables in each column paying special attention to
    a. Type of variable (categorical/numerical) and what are the units (for the numerical only)

    b. For the ones that are numerical study whether they have outliers. There is no definition for what an outlier so we can define an outlier as any observation with a value that is more than 4 times its standard deviation.

The variable that will focus our study is y and it indicates whether the household actually decided to join the bank. We will see how we can use the predictive modeling techniques seen in class to improve the efficiency making marketing phone calls.

3. Create a corr-plot using the package corrplot. You will have to install it using the command `install.packages()`

4. Run the following command `lm(y~.,data=dta_bank)`

    a. Write the structural equation that R is estimating?
    b. Comment the results.
        i. Best time to perform telemarketing tasks?
        ii. Best income groups?
        iii. Potential concerns of omitted variable Bias

Predictive Modeling and Tuning

This is a predictive modeling exercise and we have seen in class that we always divide the data set in `dta_bank_training`, `dta_bank_validating`, `dta_bank_test`.

1. Explain (in sentences) why and how we always do that.
2. From the point of view of the firm and given that we are running a predictive exercise, is there any variable that should not be included as X? If yes, please drop it.
3. Explain the problems of overfitting and underfitting.
4. Explain the meaning of the no free lunch theorem.
5. For the following 4 models, write their structural equations and comment:

```
lm1 = lm(y~age+factor(month),            data=????)
lm2 = lm(y~age+age^2+age^3+factor(month),data=????)
lm3 = lm(y~.,                            data=????)
lm4 = lm(y~.^2,                          data=????)
```

    a. Which one overfits more?
    b. Which one underfits more?
    c. Is the model that fits the training data the best one that has the best predictive power?
    d. Can we use a confusion matrix to analyze the problems a problem of underfitting?
    e. Which data set should we use to run these regressions?

Improving the predictive power

1. Make a visualization to inspect the relationship between the Y and each of the X that you have included in the regressions above.
    a. Does it look linear?

2. Use the other predictive methods seen in class (like NB classifiers or KNN) to check if you can improve the performance.

3. Do they make it better? Worse?

Causal Questions

1. When we study causality we always focus on the parameters multiplying the X variables instead of the predictive capacity of the model. We then give a causal interpretation to the estimated coefficients.
    a. Explain when in marketing is preferable a causal analysis to a predictive analysis.
    b. In the context of a linear regression, explain the concepts of a biased estimated.
2. Which of the variables could be interesting to analyze from a causal point of view. Give examples.
3. For those variables what would be the potential omitted variables problem?