**Project: Bank Marketing Data**

* [The Task](file:///C:\Ismail\Cours\UM6P\course\Madhavan\Madhavan%20Mukund_project.html#task)
* [Solving the Task](file:///C:\Ismail\Cours\UM6P\course\Madhavan\Madhavan%20Mukund_project.html#solve)

**The Task**

The "Bank Marketing Data Set" from the UCI Machine Learning Repository is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.

The classification goal is to predict if the client will subscribe a term deposit (variable y). You can find a description of the attributes at the original UCI URL, <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>.

The UCI page contains multiple versions of the data, so the version that you need to work with is here: “bank-additional-full.csv” dataset

* Download : https://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank-additional.zip

**Solving the Task**

* Do exploratory data analysis :
  + Load the data in Pandas Frame
  + Display the head and the tail of the dataset
  + Variable Identification:
    - Define the type of every variable (numerical or categorical…) and its role in the dataset (input variable or an output variable),
  + Statistics over each numerical attribute (mean, quartile, min, max, …)
  + Univariate Analysis:
    - For numerical variables: build histograms and boxplots for each numerical variable independently. These figures would give us an understanding about the variables’ central tendencies and spread.
    - For categorical variables: build a bar chart visualization that shows the frequencies in each category.
  + Bi-Variable Analysis:
    - Continuous & Continuous (Matrix): We can build a scatter plots in order to see how two continuous variables interact between each other.
    - Categorical & Categorical: A Stacked Column Chart is a good visualization that shows how the frequencies are spread between the two categorical variables.
    - Categorical & Continuous: building boxplots combined with swarmplots.
  + Detecting / Treating missing values: More of an art rather than a systematic approach and usually it depends to the problem in hand. However we describe: two different situations:
    - If for example we have only few missing values and they appear to be random we can just proceed with the deletion of these cases.
    - If we have many missing values, we don’t want to proceed into their deletions because that would end up in having a much smaller dataset which would influence the predictive model’s performance. In this case we would either replace the values with the median/mean/model or/and add another column that shows if the other variable has a missing value or not. In the latter case, ideally they newly added column should correlate with the output variable, thus creating a new variable that might be a good predictor.
  + Detecting / Treating outliers: Having many outliers in the dataset can harm the predictive model’s performance and thus would be nice to treat them. Also in this phase there is no systematic approach to deal with them. Some ideas are :
    - During the detection phase, one of the best visualization approaches to use are boxplots for univariate analysis and scatterplots for bi-variate analysis.
    - During the treatment phase, we can either delete them if they are very few or if not, we can use a special treatment like imputing them or just treat them independently by having their own predictive model.
  + Feature Engineering: During this phase we try to infer better variables/predictors out of the existing variables. Like imagine we have a date variable, we can create other new variables out of it like weekday/weekend, Monday/Tuesday…., and so on. This newly created variable can turn out to be a good predictor if it correlates somehow with the output variable.

Documentations

* Statistics of the data
  + <https://en.wikipedia.org/wiki/Interquartile_range#/media/File:Boxplot_vs_PDF.svg>
  + <https://commons.wikimedia.org/wiki/File:Normal_Distribution_Sigma.svg>
* Variable analysis
  + <https://www.formpl.us/blog/nominal-ordinal-data#:~:text=Nominal%20and%20ordinal%20data%20are,an%20interval%20and%20ratio%20data.&text=This%20classification%20is%20based%20on,does%20not%20have%20a%20number.>
* Missing data
  + <https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4>