20180415_Batch 40_CSE 7302c_CUTe

Problem Statement:

A healthcare organization together with a couple of government hospitals in a city has collected information about the vitals that would reveal if the person might have a coronary heart disease in the next ten years or not. This study is useful in early identification of disease and have medical intervention if necessary. This would help not only in improving the health conditions but also the economy as it has been identified that health performance and economic performance are interlinked. Given the data, we need to develop appropriate models to identify/predict if the person likely to have heart disease or not

About the Data:

1. The total number of records is 34,281 with 24 independent attributes and the 25rd column is the target which needs to be predicted.
2. The variables are masked and we get little information from their names.
3. Missing values are represented as NA in some columns and as -99 in some other columns

Analysis and report:

1. Analysis is expected in R
2. Prepare a separate document mentioning your experiments
3. Generate any visualizations that provide insights about the data

Hints for solving the problem:

1. ***Data Preprocessing***
   a. ***How are you to planning to work with missing values. First, different representations must be brought to same representation and then you may choose one of the imputation methods/ choose to eliminate the records. In either case justify your actions***
   b. ***In the data some of the variables may not be relevant for analysis. How such variables can be identified, and this issue be resolved. There are functions that can identify such columns.***
   c. ***Since you are expected to build two classification models, the data preprocessing for each model might be different hence prepare the data accordingly***
2. ***Exploratory data analysis:***
   a. ***This is where we would like to see, how well you have understood the data. Any insights from the data given at this stage might fetch you good scores***
3. ***While running glm, you might get a warning that fitted probabilities numerically 0 or 1 occurred. Report/ Analyze why such warning might have occurred and how do you resolve it.***
4. ***Observe the ratio of positive to negative classes in the data (1 being positive and 0 is negative). Do you think the number of positive instances sufficient?. In such cases we can***

synthesize new samples of a class (this is statistically approved technique). The method is known as SMOTING. Would you want to try it?