Data Mining - CA 2

This CA if focused on two-stage Missing Data Imputation.

In the dataset provided, the first column is labelled as “Income” and is the column to be imputed. Other variables are the potential predictors in the models explained below.

The first step is to determine the missing values of “Income” that are actually equal to 0. To do this, you are required to define a new variable, say “Binary_Income” with 0, 1 and missing values, for the cases which have 0, non-zero and missing values in the “Income” variable, respectively. Then use the cases with “Binary_Income” equal to either 0 or 1, to develop a logistic regression model. The outcome of this model would allow you to predict the missing values in the “Binary_Income” variable, as 0 or 1. For any case which “Binary_Income” is predicted as 0, put the actual “Income” equal to 0 as well.

Those cases who their “Binary_Income” is predicted as 1, are used along with other cases with non-zero “Income” to develop a linear regression model. The outcome of such model would allow you to predict the missing values for the “Income” variable, for the cases which their “Binary_Income” variable is predicted as 1 and therefore are expected to have non-zero “Income”.

Report:

- The results for both logistic and linear regression models.
- The number of cases with predicted “Binary_Income” as either 0 or 1.
- The Mean and SD for the non-zero predicted “Income”.

Deadline: 25 April 2020