Statistical Modelling & Machine Learning HW3
(Due: 11/22/2020, Sunday)

Instruction:

- There is no correct or unique answer in this homework.
- I will give your HW score based on your results and analysis procedure.
- Submit your HW solution with a `R` code file. Also, your solution should include the **brief description of your analysis procedure**.
- Your `R` code should show the procedure that you obtain the final result (**Do NOT include `R` codes for all procedure that you have tried**).

1. Consider the `train` and `test` data files. Using the `train` dataset, build your prediction model and then apply your prediction model to `test` dataset and compute the test misclassification rate.

Description of data:

- The data have information on credit card customers. The goal of analysis is to accurately predict customer's payment default status based on their demographic information and credit card history.

Description of variables:

- `X1`: Amount of the given credit.
- `X2`: Gender (1 = male; 2 = female).
- `X3`: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- `X4`: Marital status (1 = married; 2 = single; 3 = others).
- `X5`: Age.
- `X6 - X11`: Monthly payment records from April to September [i.e., `X6`: Payment record for April;...; `X11`: Payment record for September] (-2 = pay two months early; -1 = pay a month early; 0 = pay in correct time; 1 = payment delay for one month; 2 = payment delay for two months;...; 8 = payment delay for eight months; 9 = payment delay for nine months and above).
- `X12 - X17`: Amount of bill statement from April to September [i.e., `X12` = amount of bill statement in April;...;`X17`: amount of bill statement in September].
- `X18 - X23`: Amount of the previous payment [i.e., `X18` = amount paid in April;...;`X23`: amount paid in September].
- `Y`: Default of payment (1 = yes; 0 = no).

Instruction and suggestion for analysis:

- You can use any prediction models including data models and algorithmic models
- There are missing values in the `train` dataset. You might need to impute the missing values.
- When you impute missing values or build your model, you cannot use the `test` dataset. The `test` dataset should be used only for calculating the test error rate.
- You might need to create new input variables and/or transform the variables.
- Also, you might need to filter or select input variables.