CSC498H/BIF524 Data Mining Fall 2021

Assignment -1-

General guidelines: This is a graded assignment on which you need to work individually. Use R Markdown to show your work in detail. The deadline is strictly on Friday, October 1st, at 11:45pm.

Task I: Data Preprocessing

Suppose that you want to build a model to predict whether a client has a health coverage insurance. You collected a set of information that you think can help predict the probability of health coverage.

1- Load the "Clients" dataset into R. What does the collected information on each client consist of?

2- Using the *summary* function, comment on *age*, *income*, and *housingStatus*.

3- Based on a subset of the data with $0 < age < 100$ and $income > 0$, is there a correlation between *age* and *income*? Plot both variables against each other and comment on the variations.

4- What is the number of missing values in *housingStatus*, *recentlyChangedHousing*, and *numberCars* variables? Find out whether those missing values come from the same observations? Comment on how to deal with them accordingly.

5- Now check the variable *works*. What can you say about the number of missing values? Does it make sense to remove all observations with NAs? If you think about the possible meaning of NA for this variable, it seems more reasonable to create a new variable *fixedWorks* with an additional level of value "missing" for those observations, while keeping "employed" and "not employed" for the two other possible levels. Find a way to define such variable in R – Hint: use *ifelse*.

6- What is the type of variable *income*? What is the number of missing values in this variable? Assuming that the observations with missing *income* values have the same distribution as clients with specified *income* values, create a new variable *fixedIncome* in which you fill the mean income value in place of missing *income* values – Hint: you can also use *ifelse*.

7- Recall that you want to use this dataset to predict whether a client has a health coverage insurance based on collected information about him/her. Do you think that the actual value of *age* matter? Or is it the ranges of corresponding values that need to be considered to predict health coverage? In this direction, it may seem reasonable to discretize such variables, i.e. convert them from continuous to discrete variables.

Use the function *cut* to define a new variable *age.range* by dividing *age* into ranges, with breaks $0,25,65, and\ INF$.

Task II: The goal is to generate a model to predict diamond prices based on a set of features.

1- Use *ggplot* to:

a. Plot *price* against *carat*.
b. Plot *cut* against *price*.

2- Generate a simple linear regression model with *price* as response and *carat* as predictor. Comment on the model summary.

3- Generate a multiple linear regression model with *price* as response and *carat*, *clarity*, and *color* as predictors. Comment on the model summary.

4- Generate a multiple linear regression model with *price* as response and all attribute in the dataset as predictors. Comment on the model summary and compare it to the previous model.