Fall 2022
MIS 572/CM 503 Introduction to Big Data Analytics

Group Exercise 2

- Graded out of **100** points. Please typeset your answers, save as an R source code file with title "Your group ID_Exercise_2.R" (e.g. Group01_Exercise_2.R).
- Please submit your code to NSYSU Cyber University before **12/17 11:59pm**. **No late submission**.
- DO NOT use any loops in your answers. Also notice that your code must follow the suggested programming and data analysis styles discussed in the class.

1. Please load the built-in *Credit* data in the ISLR package and remove the variable "ID". Enter "?Credit" to check out the data description.

   1.1 **[10 pts]** Split the dataset into training (70%) and testing (30%) sets with random seed *set.seed(1)*. After that, rescale the training and testing sets if necessary.

   1.2 **[10 pts]** Please rank the importance of variables by absolute standardized regression coefficients of general linear models that predict "Balance".

   1.3 **[10 pts]** Please create a leave-one-out cross-validation (LOOCV) MAE to evaluate general linear models.

   1.4. **[25 pts]** Please build general linear models and perform the forward selection discussed in the class. We here consider the most important variable identified by the ranking we created in 1.3 as the "force-in" variable. Then, add the remaining variables to the baseline model with the force-in variable, respectively. Please report the LOOCV MAEs in the first round of the forward selection process. Note that there are *p* (the number of predictors) training/validation MAEs needed to report in total: the baseline model (y ~ x1) with *p-1* models (y ~ x1 + x2, y ~ x1 + x3,..., and so on). After the first round of the forward selection, what is the best linear model in terms of the lowest validation error?

   1.5. **[20 pts]** What is your best linear model in terms of the testing MAE? Please report and plot training, validation, and testing MAEs of your model selection process(at least 5 model errors). You may use any model/feature selection technique (e.g., forward selection, adding more interaction terms, and so on) discussed in the class. Does your model overfit or underfit the training data? And why?

2. Please load the given dataset "mushrooms.csv".

   2.1 **[5 pts]** Convert the variable type from character to factor and replace the variable names with the following "col_name".

```
col_name <-
c("class", "cap_shape", "cap_surface", "cap_color", "bruises", "odor", "gill_attachment", "gill_spacing", "gill_size", "gill_color", "stalk_shape", "stalk_root", "stalk_surface_above_ring","stalk_surface_below_ring", "stalk_color_above_ring", "stalk_color_below_ring", "veil_type", "veil_color", "ring_number", "ring_type", "spore_print_color", "population", "habitat")
```

2.2 **[5 pts]** Refer to the given data description file “Mushrooms_Readme.pdf” for more information about the data. Please remove those records with missing “stalk_root”.

2.3 **[5 pts]** Please fit a simple logistic regression model and computer the predicted probability of class = poisonous when mushrooms have bruises (bruises = t) or not (bruises = f).

2.4 **[10 pts]** Create a crosstab "bruises by class" and calculate the percentages. Then compute odds ratio (OR) using this crosstab and briefly describe your finding.