Fall 2023
MIS 413_572/CM 503 Introduction to Big Data Analytics

Homework 1

- Graded out of **150** points. Please typeset your homework, save as an R or Python source code file with title "your student ID_Homework_1" (e.g. B104020001_Homework_1.R or B104020001_Homework_1.ipynb).
- Please submit your code to NSYSU Cyber University before **11/26 11:59pm**. **No late submission**.
- **Do not use any loops or repeat the same code multiple times as a replacement for loops** in either R or Python code. Ensure code follows the programming and data analysis styles discussed in class by including comments explaining each section:
    - **5** points will be deducted from code without comments in each part.
    - **5** points will be deducted if loops or repeats are used in each part.

---

1. Please load the given "Dengue_y2008_2018.csv". The datasets contain the responses of dengue fever infections data each month from 2008 to 2018 in Bangladesh. "DENGUE" is the target variable, and the other variables are predictors. Consider the following questions. (70%)

   YEAR - observation year (2008 - 2019)
   MONTH - observation month (1 - 12)
   MIN - average minimum air temperature of corresponding month (10.6 - 26.5)
   MAX - average maximum air temperature of corresponding month (23.5 - 35.8)
   HUMIDITY - average relative humidity in % of corresponding month (67.5 - 88.4)
   RAINFALL - average rainfall in mm (0 - 689)
   DENGUE - number of dengue incidents (0 - 3087)

   1.1. **[5 pts]** Assume that we have received the new data for the year 2019 (Dengue_y2019_.csv). Please load the new dataset and merge it with the data from other years (Dengue_y2008_2018.csv).

   1.2. **[5 pts]** Check the dataset for any missing values (NAs). If there are any observations containing NAs, display the entire observation and then remove it.

   1.3. **[5 pts]** Are there any duplicate observations? If any duplicates exist, keep only one instance of the duplicated observation and remove the rest.

   1.4. **[10 pts]** Create a new column called "SEASON", where Dec.-Feb. is winter, Mar.-May is spring, Jun.-Aug. is summer, and Sep.-Nov. is fall. Then, group the data by season to sum the "DENGUE" values.

   1.5. **[5 pts]** Convert columns "YEAR", "MONTH" and "SEASON" to factor in R and categorical in Python.

   1.6. **[10 pts]** For continuous variables, create density plots to understand the distribution of the data.

   1.7. **[10 pts]** Perform a series of bivariate analysis to check whether the continuous

variables are associated with the “DENGUE”.

1.8. **[5 pts]** Write a function that computes Mean Absolute Error (MAE), which is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right|$$

1.9. **[5 pts]** Please split the dataset into a training set (80%) and testing set (20%) with random seed = 1.

1.10. **[10 pts]** Build linear models (at least two) with the training set, then report the training and testing MAEs (round up to the fourth decimal digits), in terms of MAEs, which model performs better? Explain your answer.

2. Please load the given “Diamonds.csv”. This dataset contains the prices and other attributes of almost 54,000 diamonds. “PRICE” is the target variable, and the other variables are predictors. Consider the following questions. (60%)

price - price in US dollars ($326 - $18,823)
carat - weight of the diamond (0.2 - 5.01)
cut - quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color - diamond colour, from J (worst) to D (best)
clarity - a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x - length in mm (0 - 10.74)
y - width in mm (0 - 58.9)
z - depth in mm (0 - 31.8)
depth - total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43 - 79)
table - width of top of diamond relative to widest point (43 - 95)

2.1. **[5 pts]** Show parts of the dataset and browse it, you may notice some ambiguous columns, like “Unnamed: 0” and “x”, “y”, “z”. Use any method to remove the extraneous column “Unnamed: 0” and give the useful columns “x”, “y”, “z” meaningful names.

2.2. **[10 pts]** Convert any character columns to factor in R and categorical in Python. Then, fit a log transformation to the target “price”.

2.3. **[10 pts]** Consider a series of bivariate analyses on "price" vs. the rest variables. Specifically, plot your data and perform bivariate statistical tests to understand the relationships among the variables. Are “carat” and “cut” associated with “price”? Use any statistical methods to justify your answers. Also notice that you may consider any data transformation on the “price” that helps understand the associations or better predict the “price”.

2.4. **[10 pts]** Please split the dataset into a training set (80%) and testing set (20%) with random seed = 1. Then rescale continuous variables into the values ranging from 0 to 1 without centralizing.

2.5. **[5 pts]** Build a linear model with the rescaled training set, then report the training and testing MAEs (round up to the fourth decimal digits).

~~2.6. **[10 pts]** Remove the predictors with higher p-values(> 0.05), then build a new linear model. Does the new model have lower errors in terms of training and testing MAE? Explain why the new model has good/bad performance.~~(上題2.5所有p-value<0.05,

故本題送分)

2.7. **[10 pts]** Again, we would like another new model that considers all the two-way interactions without removing any predictors. Please report the training and testing MAEs. Does the new model have lower errors in terms of training and testing MAE? Can this complex model with more parameters improve the prediction?

3. **[20 pts]** Considering the linear regression models from question 1 and 2, which dataset do you think is well-suited for linear regression modeling and which is not? For any datasets that are not well-suited, explain why linear regression performs poorly. How could the prediction accuracy be improved for those cases? Provide detailed explanations to support your answers.

-- End --