Fall 2023

MIS 413_572/CM 503 Introduction to Big Data Analytics

Group Exercise 2

- Graded out of **100** points. Please typeset your homework, save as an R or Python source code file with title "your student ID_Exercise_2" (e.g. B024020001_Exercise_2.R or B024020001_Exercise_2.ipynb).
- Please submit your code to NSYSU Cyber University before **12/31 11:59pm**. **No late submission**.
- Notice that your code must follow the suggested programming and data analysis styles discussed in the class.

1. Please load the given Student Alcohol Consumption data *student-por.csv*. Check out more details about the data on Kaggle.

   1.1. **[10 pts]**
   Please check and remove records with any NAs, and drop the grade columns G2 and G3. Then convert all the categorical variables into R factors or Python Pandas Categoricals.

   1.2. **[10 pts]**
   Here we consider the first period grade (G1) as the outcome/target. Please draw the density plot of G1 and perform normality tests to check if it is approximately normally distributed. Then, perform a proper bivariate test to check whether sex is associated with G1 (the significant level is 0.05).

   1.3. **[5 pts]**
   Please create a function rmse(y_true, y_pred), which computes the RMSE of the model prediction.

   1.4. **[25 pts]**
   Please split the data into training (70%) and testing (30%) sets with a random seed 0, and train models to predict the outcome/target G1. (Note that you should rescale the data if needed). Use any statistical learning and feature selection techniques to create a better model with low testing RMSE. Please report both the training and the testing RMSE of your models. Note that your testing RMSE must be at least lower than 2.35

2. Please load the given data Heart Disease Health Indicators Dataset (heart_disease_health_indicators_BRFSS2015.csv).

2.1. **[15 pts]**
Split the data into training (70%) and testing (30%) datasets with random seed 0). Fit Logistic regression models that predict “HeartDiseaseorAttack”.Use or create any variables that may better predict the HeartDiseaseorAttack. What are the accuracy of predictions on both the training and the testing datasets, given the default probability cutoff value 0.5?

2.2. **[15 pts]**
We can see that there is a class imbalance problem with the outcome/target (HeartDiseaseorAttack). We understand that adjusting predicted class probability cutoff may help predict the rare cases. What is the optimal cutoff value based on Youden's J index? Please also report your model True Positive Rates (Sensitivities) with different cutoff values (0.5 and the “optimal” value).

2.3. **[20 pts]**
Plot the ROC curves of your models for both training and testing datasets. Compare and report your model performance in terms of AUCs.