Cardio Vascular Disease

Framingham Heart Study
A Project of the National Heart, Lung, and Blood Institute and Boston University

Data Set Information:

- Committed to identifying common factors contributing to cardiovascular disease (CVD).
- Setup in the town of Framingham, MA in 1948.
- Random sample consisting of 2/3rds of adult population in the town.
- 5209 men and women participated.
- Age range: 30-62
- People who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke.
- Careful monitoring of Framingham Study population has led to identification of major CVD risk factors.
- Led to development of Framingham Risk Score, a gender specific algorithm used to estimate the 10-year cardiovascular risk of an individual:

http://cvdrisk.nhlbi.nih.gov/

Attribute Information:

4240 observations; 15 predictor and 1 predicted variables

• *TenYearCHD* – To be predicted. Risk of having a heart attack or stroke in the next 10 years.

Predictors

• Demographic Risk Factors

– male: Gender of subject – Yes or No

– age: Age of subject at first examination

– education: some high school (1), high school (2), some college/vocational college (3), college (4)

• Behavioural Risk Factors

– currentSmoker: Yes or No

– cigsPerDay: No. of cigarettes smoked per day if smoker

• Medical History Risk Factors

– BPmeds: On BP medication at the time of first examination – Yes or No

– prevalentStroke: Did the subject have a previous stroke – Yes or No

– prevalentHyp: Is the subject currently hypertensive – Yes or No

– diabetes: Does the subject currently have diabetes – Yes or No

• Risk Factors from First Examination

– totChol: Total cholesterol (mg/dL)

– sysBP: Systolic blood pressure (the higher number in BP result)

– diaBP: Diastolic blood pressure (the lower number in BP result)

– BMI: Body Mass Index (kg/m2)

– heartRate: # of beats per minute

– glucose: Blood glucose level (mg/dL)

Task -1

Please kind the following task regarding the dataset:

1. Read the dataset
2. Identify **categorical** and **numerical** variables in the dataset and write in *markdown.*
3. **Use all recommended plots will be add-on**
4. Create Dummy Variable for Categorical Data **(n-1 columns)**
5. Split the data into training and testing set (70 % training and 30 % training)
6. Build Logistic regression model with training data.
7. Take default threshold p = 0.5
8. Interpret the results
9. Compute **Confusion Matrix** for training set and testing set
10. From confusion matrix compute
    a. Sensitivity, Specificity, Precession and Accuracy for training set
    b. Sensitivity, Specificity, Precession and Accuracy for testing set
11. Also compute **Kappa Score for** training and testing set (optional)
12. Draw ROC curve Suggest the approx. threshold value for probability of success
13. Plot logistics regression for best threshold probability value which you feel
    a. Compute Confusion matrix
        i. Training data
        ii. Testing data
    b. Computer Sensitivity, Specificity, Precession and Accuracy for both training and testing confusion matrix.
14. Any additional recommendation want to made in this model is add-on

Submit the **jupyter notebook** file and also **pdf** of jupyter files.