Individual Coursework Submission Form

Specialist Masters Programme

Surname: Heijliger-Krogulski	**First Name: Boris**
MSc in: Actuarial Science with Business Analytics	**Student ID number: 230063332**
Module Code: SMM636	
Module Title: Machine Learning	
Lecturer: Dr Rui Zhu	**Submission Date: 22/03/2024**

Declaration:
By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work, I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

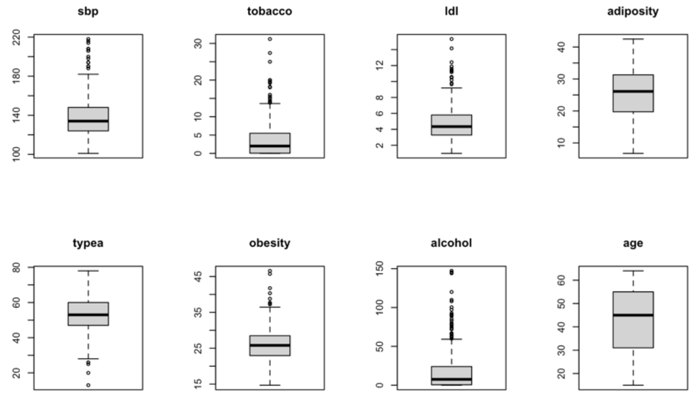
Marker's Comments (if not being marked on-line):

Deduction for Late Submission: □

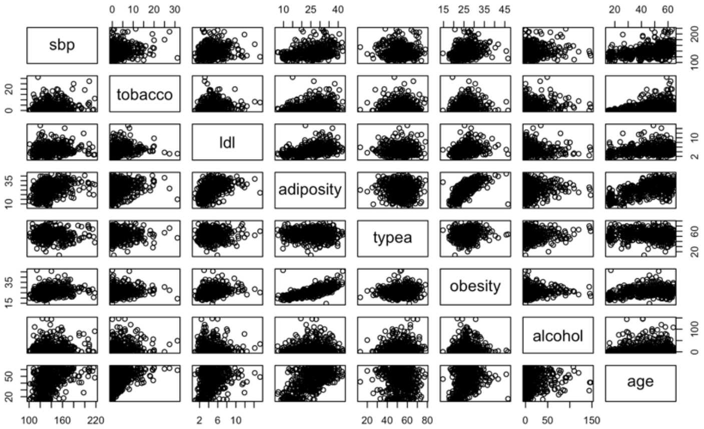
Final Mark: %

Question 1:

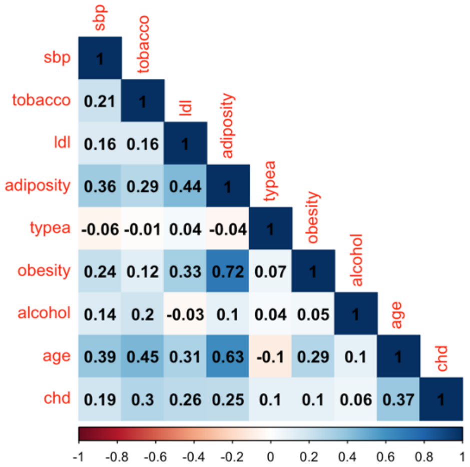
Prerequisites include loading multiple libraries such as '*tidyverse', 'ggplot2', 'caret',* and others to use functions from these packages. Then, loading the dataset, we can attach it to call variables easily. Starting with a summary, we can see some basic statistics for each of the variables from the dataset. This isn't indicative of much, so we create boxplots to easily visualise the data:



For a number of these health factors such as '*sbp', 'tobacco', 'ldl', 'obesity',* and '*alcohol',* we have many outliers above the maximum. This shows that in our sample, some patients have unusually high levels of cholesterol, smoke more tobacco, and drink more alcohol than most others. We will now check correlation between variables, and how they interact with each other. This is done by taking the numeric variables, taking out *'chd'* as it's our response variable, and plotting a '*pairs()'* function to gain a general understanding of the relation. we will no longer include the '*famhist'* variable in our dataset. The rationale behind this is that, by the American Heart Association, while family history might influence the development of CHD, it is only significant for first-degree relatives who have experienced heart complications before the age of 55 for males and 65 for females. Therefore, its applicability in assessing risk is limited. Since we are not given this data, including this variable may lead to inaccurate results, hence we leave this out.



Whilst this shows some useful information, it is too cluttered to see exactly what relations are between each of the variables. To counteract this, we can use a correlation plot to condense the information down, and make some interesting findings, making sure to include '*chd*'.



Some correlations make sense, like the strong positive between obesity and adiposity, age and adiposity, and age with tobacco use, whereas some are more perplexing, like how alcohol is weakly correlated with everything. In terms of CHD, the most correlated factors are age, tobacco, LDL, adiposity, and SBP, whereas alcohol, obesity, and type-a behaviour are weakly related. This gives us a good initial understanding of what the biggest risk factors are.

Question 2:

To fit a logistic regression, we must split the sample into predictor and response variables. Since we are modelling coronary heart disease, '*chd'* will be the response variable, and we take all other variables as predictors. We will create an initial model using a generalised linear model (GLM) to use all the predictors as a function of the response variable. We use the binomial family and select a penalty (α) of 0. We then create a similar model, named '*final_model'*, using the same penalty term, yet only using the lowest λ from our initial model. Outputting the coefficients of this model shows that:

1. **Intercept (s0)**: represents the estimated log-odds of the response variable (CHD) when all predictor variables are zero. In this case, it's approximately **-4.78**. Since the log-odds are negative, this suggests a lower probability of coronary heart disease (CHD) when all other predictors are zero.
2. **sbp:** For every unit increase in systolic blood pressure, the log-odds of CHD increase by approximately **0.0032**.
3. **tobacco**: For every unit increase in tobacco usage, the log-odds increase by approximately **0.0338**, indicating that consumption is positively associated with CHD risk.
4. **ldl**: A one-unit increase in LDL corresponds to an increase in the log-odds by approximately **0.0698**.
5. **adiposity**: For every unit increase, the log-odds of CHD increase by approximately **0.0104**.
6. **typea**: A one-unit increase in Type A behaviour leads to an increase in the log-odds of CHD by approximately **0.0121**.
7. **obesity**: A one-unit increase in obesity results in a decrease in the log-odds of CHD by approximately **-0.0105**. Surprisingly, this coefficient is negative, suggesting that higher obesity levels are associated with a slightly lower CHD risk.
8. **alcohol**: For every unit increase in alcohol consumption, the log-odds of CHD increase by approximately **0.0009**.
9. **age**: A one-year increase in age corresponds to an increase in the log-odds of CHD by approximately **0.0165**.

10. **chd**: The coefficient for the response variable itself is approximately **4.189**. This indicates that the presence of CHD significantly increases the log-odds of CHD (which is trivial, as it's a direct predictor).

Using this model, we will generate predicted probabilities for each patient, and define a threshold of 0.5. This will allow us to classify based on whether their predicted probabilities are greater than or equal to this. If they are (1), then we classify them as having CHD, otherwise not (0):

0	1
302	160

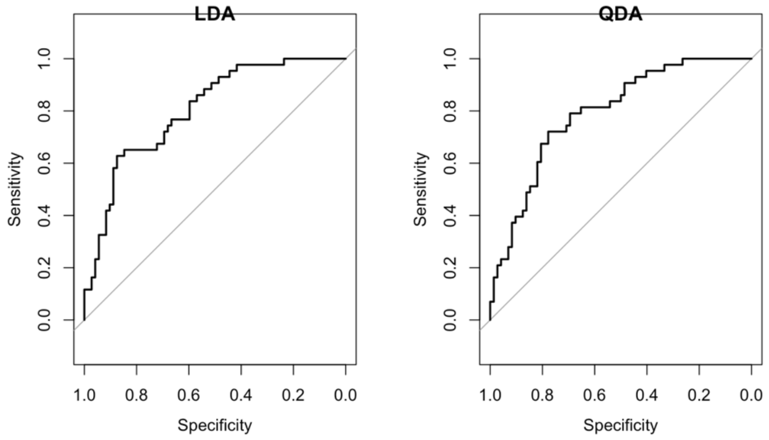
Therefore, roughly 35% of patients are classified as having CHD.

Question 3:

We will explore two other classifiers, namely Linear/Quadratic discriminant analysis (LDA/QDA). Firstly, we split into 75% testing and 25% training data. We apply the models with the same response variables. Now we can compare efficiency between them, starting with the confusion matrix.

LDA	0 (predicted)	1 (predicted)	QDA	0 (predicted)	1 (predicted)
0 (actual)	64	8	0 (actual)	62	10
1 (actual)	22	21	1 (actual)	24	19

LDA has a higher number of true positive and true negative results, and a lower number of false positive and false negative results. This information initially suggests that the LDA model has a higher accuracy than that of QDA. Now we will compare ROC curves.



Classifiers that have a curve close to (1,1) indicate a stronger performance, which looks like QDA, but to be sure, we will analyse the AUC for these curves. The AUC goes from 0 to 1 (as a percentage), measuring the ability of a classifier to distinguish positive and negative instances. For LDA, we get an AUC of 80.3617% and for QDA, the AUC is 79.167%. Whilst these values are incredibly similar, we are better off using LDA to classify our patients, as it has beaten the QDA classifier. In terms of results, we obtain:

0	1
86	29

So out of our testing data (25% of the whole dataset) we classify roughly 25% of patients as having CHD.