

**BC2406 CBA**

Name: Venkat Subramanian

Matriculation Number: U1921075D

Email : VENKAT001@e.ntu.edu.sg

Table of Contents

**Part A2**

Q1.1 Data Quality Issue2

Q1.2 Data Exploration3

Q2 Summary Table 6

**Part B7**

Q3.1 CART Model7

Q3.2 Logistic Model8

Q3.2 Logistic Model9

Q4 Model selection for Hospital10

Q5 Key Information for Hospital11

**Part C12**

Q6 Missing Information in CP table 12

Q7 Comments on the approach taken by the paper 13

**Part A**

Q1.1 – Data Quality

There are mainly 2 Data Quality issues encountered in the dataset. One of them was the presence of categorical variables that have numeric data. Even after reading the csv, with stringAsFactors as True, there were still many categorical variables that were classified as numeric. These variables include Sex, Fbs, ExAng and slope. Before exploring the data, these variables were converted to categorical using the factor function.

The other data quality issue face was regarding the presence of NAs in the data. This wouldn’t pose much of a problem for Classification and Regression Tree (CART) modelling since it is able to make use of surrogates. But the presence of NAs will affect the result of the logistic regression (more information will be covered in Part B, Q3.2)

Due to the lack of information regarding a person’s maximum and minimum RestBP, Cholesterol, Ca, MaxHR and Oldpeak, an assumption that the data provided for these variables are accurate had to be made. By looking at the summary of the dataset (Fig 1), it is observed that the data regarding the Age does not require additional data cleaning. This is due to the fact that the maximum Age of the data is still lesser than the oldest person in the world who is around 100 years old and the minimum age is of that of a working adult, therefore no data cleaning required.

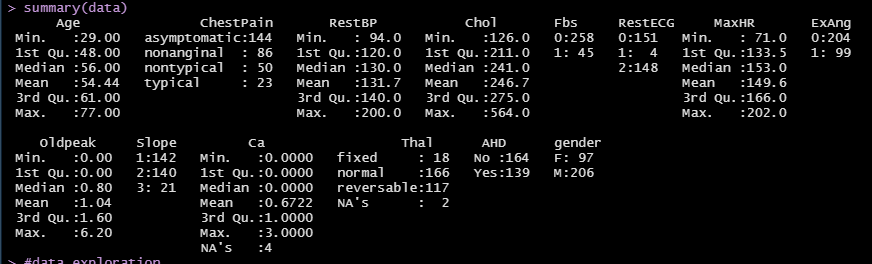


Figure :Summary of the Dataset

The variable Ca was not converted to be a categorical variable, since the numeric data does not have any special meaning behind unlike Fbs or AHD.

Even though a person’s MaxHR can be calculated by subtracting their age from 220[[1]](#footnote-1), these patients might have other health complication that might have resulted in a lower or higher MaxHR than the calculated MaxHR. Even though the minimum value of MaxHR was 71(implying that the age of the person was 149) no cleaning was done on this variable.

For easier visualization of data, a new column gender was introduced. Gender has the same values of Sex but instead of 0 and 1, gender contains F and M to represent Females and Males respectively.

**Part A**

Q1.2 -Data Exploration

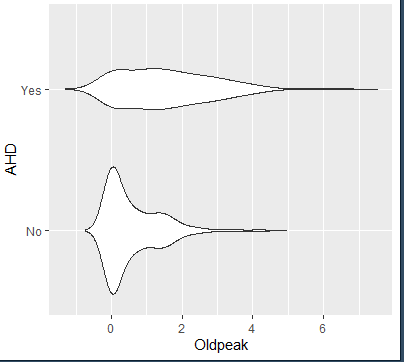
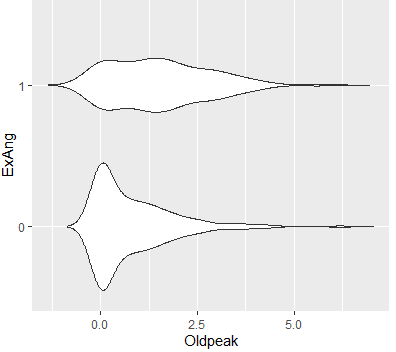
By conducting data exploration on the individual variables, there were few interesting findings found. One of them was regarding the effect of Oldpeak on AHD. Since Oldpeak refers to the ST Depression Induced by Exercise Relative to Rest, it might have been affected to the level of exercise as well. When exploring this relationship, it was found that the distribution of Oldpeak against AHD (figure 2) is almost the same as the distribution of Oldpeak against ExAng. This implies that Oldpeak affects the ExAng (or vice versa) which then affects AHD.

Figure 3: Oldpeak against ExAng

Figure 2: Oldpeak Against AHD

Therefore, when Oldpeak is lesser than 2, ExAng is more likely to be 0, therefore AHD is most likely a “No”. Since Oldpeak might have been affected ExAng (Exercised Induced Angina), the level of Chol might have been affected by Oldpeak as well since cholesterol levels usually decrease with exercise.

Chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 4: Oldpeak against Chol

Figure 5: Smooth plot for Oldpeak against Chol

From Figure 4 we can see that there is generally a positive correlation between Oldpeak and Chol. However, it is observable that for AHD = “Yes” there is a negative correlation whereas when AHD = “No” there is a positive correlation. This relationship is also shown in Figure 5. This implies that instead of just one variable affecting AHD, AHD might have been caused by both Oldpeak and Chol. Therefore, a new variable Oldpeak\_per\_Chol was introduced (explained further in Q3)

**Part A**

Q1.2 -Data Exploration (continuation)

Chart, bar chart

Description automatically generatedA picture containing chart

Description automatically generatedThe effect on Age on AHD was also explored. From figure 6, it is observed that most of the patients that are between 55 to 60 of Age tend to have AHD. Also, it was observed that most of the females tend to be within this Age range (Figure 7). Males tend to have a higher probability to get AHD than females. Even though females tend to be within 55-60 , females tend to not have AHD (Figure 8). Further exploration was conducted to find out possible reasons as to why females have lesser cases of AHD

Figure 8: Gender Against AHD

Figure 7: Age against gender

Figure 6: Age against AHD

Chart

Description automatically generated

From further data exploration, it was found that females tend to have non-anginal and non-typical chest pains (Figure 9). And these types of chest pain tend to be the symptoms of a non-Angiographic Heart Disease (Figure 10). Even though this might explain why females do not have AHD, I believe that the chest pain experienced is due to the effect of AHD. Thus, it does not still quite explain the reason as to why females do not have AHD.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure 9: ChestPain against Gender

Figure 10: Chestpain against AHD

It seems that female tend to have Thal as normal (Figure 11), and from the univariate analysis of Thal it was found that AHD tends to be a “No” when Thal is normal (Figure 12). Furthermore, by seeing the trend of ChestPain against Thal, it was observed that when Thal is normal majority of the patients experience non-anginal and non-typical chest pains (Figure 13). This falls in line with our prior observation that females tend to have non-anginal and non-typical chest pains and that these types of pain tend to be the symptom of patients without AHD. Therefore, Thal could be the reason Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedas to why females tend to have lesser cases of AHD.

Figure 13: ChestPain against Thal

Figure 11: Thal against gender

Chart, bar chart

Description automatically generated

Figure 12: Thal against AHD

Another interesting finding during data exploration was that even though Age affects most of the variables to have AHD, Ca might have a larger effect on AHD than Age. This is because when a scatter plot of Age against Ca was obtained (Figure 14), it was observed than when Age is greater than 55 and when Ca is 0 majority of the cases were “No” cases. But in the same Age range, when the Ca values were increased it was noticed that more of the case were “Yes”. Therefore, Ca might be a more influential factor than the Age.

Chart

Description automatically generated

Figure 14: Age against Ca with gender on AHD

**Part A**

Q2- Summary Table and its explanation

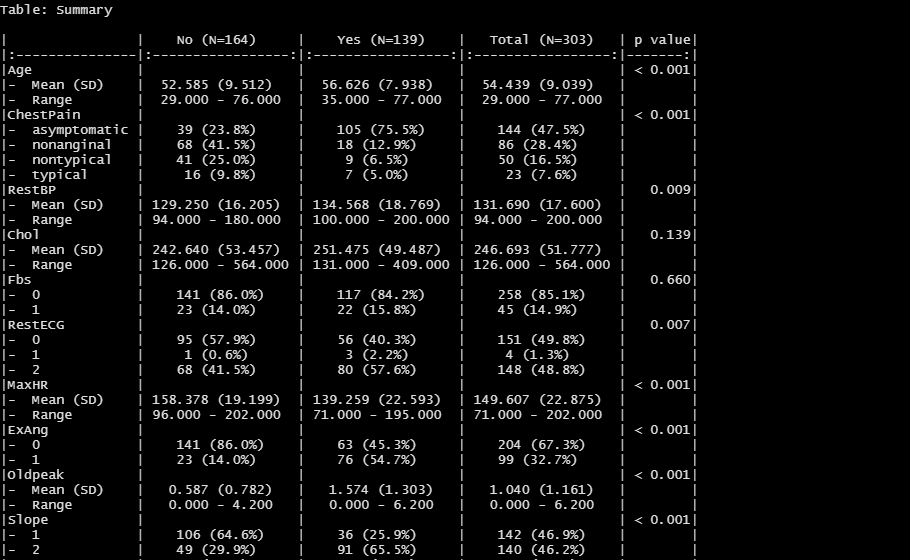
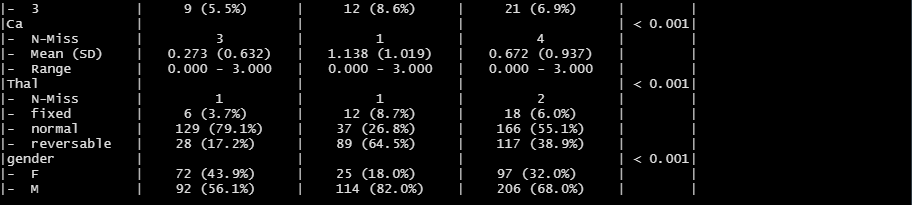


Figure 15: Summary of all variables

From Figure 15 we can obtain important information such as the mean, standard deviation and the range for continuous variable for each AHD type. And for categorical variable we are able to obtain the number of cases and the percentage of that specific category. For example, from this table we can detect that the mean age of those patients who had AHD is higher than those who do not, which was what was observed in the Q1 as well. An example of categorical variable is ChestPain, from the summary table it can be noted that for non-AHD cases it tends to be nonanginal and for AHD cases the ChestPain tends to be Asymptomatic, similar to the result observed in Q1. Furthermore, we can observe the trend of the entire dataset from focusing on the Total column .In addition to that we are also able to get the number of NAs in each of the variable in each of the outcome category.

Some of the data that we can get from Figure 15 is that, those that have AHD tend to be older males that have asymptomatic ChestPain, higher BP, higher Chol, lower MaxHR, have a ExAng of 1, higher Oldpeak , have a slope of 2 and have more Ca . It can also be observed that the Fbs doesn’t affect much since both categories of AHD have equal cases of Fbs.

**Part B**

Q3.1- CART Model

Since CART model can make use of surrogates, the CART model was trained on the dataset with the NAs and the additional variable Oldpeak\_per\_Chol mentioned in Q1.2. The Oldpeak\_per\_Chol is obtained by dividing the Oldpeak with its cholesterol level. Before training the CART model, the dataset was split into 70% train and 30 % test. The data were stratified on AHD, and since the number of Yes and No cases are the approximately the same there is no need to create a new balanced dataset to train the model. After growing the tree to its max, the optimal prune was found to be 0.045 (Figure 16). After pruning the tree, the optimal CART model was used to predict the outcome of the test set.

Chart, line chart

Description automatically generated

Figure 16: Cp values against cross validate relative error and number of splits

Timeline

Description automatically generatedThe resulted CART Model is shown in Figure 17 and the result of testing the model on the test set is shown in Figure 18. From the CART Model we can see that the newly engineered variable plays a significant role in classifying the variable. The overall accuracy of the CART Model was calculated to be 72.7%

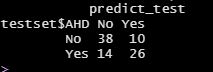


Figure 18: Result of CART Model

Figure 17: Result of CART Model

**Part B**

Q3.2- Logistic Model

Another model that can be used to predict AHD is the Logistic Regression Model. Unlike the CART model in Q3.1, Logistic Model is unable to make use of surrogate to model. Therefore, there is a need to clean the main dataset before training the model. Since there is only 6 NAs in the dataset (shown in Figure 15 in Q2) which is 1.98% of the data, these NAs can be dropped. After dropping the NAs ,the model was trained with the same trainset as CART model but without the NAs. After training the Logistic Model, it was observed that there was no influential outlier (Figure 19). Then the Logistic Model was optimized by using the step function. The step function will select those variables that reduces the total AIC value of the model. Once the optimal logistic model was obtained the outcome was predicted using the newly optimized model. Initially the threshold was set at a value of 0.5. The predicted probability was plot against the outcome variable to see if the threshold of 0.5 is suitable (Figure 20).

Chart, box and whisker chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 19: Cooks Distance for Logistic Regression before step

Figure 20: Predicted Probability against AHD

Timeline

Description automatically generatedFrom Figure 20 we can see that there is a clear distinction between the “Yes” predicted probability and the “No” predicted probability. A threshold of 0.5 might not have been suitable since it will still misclassify some of the “Yes” cases. To find a suitable threshold a CART Model was built on the predicted probability to find the best split. And after pruning it was shown that the threshold of 0.26(Figure 21) is more suitable for the logistic model. The result of the logistic regression model with the new threshold is shown in Figure 22

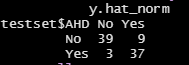
****

Figure 2: Confusion Matrix using optimal Threshold

Figure 21: Cart Model to identify optimal threshold

Q3.3- Which Model Performed Better

Due to the ability to handle missing data, CART model was trained on a train set that had missing data. But the logistic regression model was not trained on a train set that did not have any missing values. To make an unbiased decision on which model performed made, the test set was made the same. The test set did not include any missing data. If we compared on the performance of the models, then the logistic model outperformed the CART model. Since the logistic model was able to have an overall accuracy of 86.4% whereas the CART model only had 72.7 %.

**Part B**

Q4- Which model is more suitable for the Hospital

The aim of the Hospital is to identify the AHD cases accurately. To do so the model chosen must have a higher sensitivity (true positive ratio). If we took the “No” cases as negative and the “Yes” cases as positive we want a model that can predict the “Yes” cases more accurately .The sensitivity of the logistic model is 92.5%, this is calculated by the true positive predicted and dividing it by the total number of positive cases in the test set. The sensitivity of the CART Model is 65 %. Therefore, since the logistic model is able is able to predict the general cases more accurately and it has a higher chance of predicting the AHD cases correctly. Therefore, Logistic model would be preferred for the hospital.

In some situations, the logistic model is not suitable. The logistic model requires 7 variables, whereas the CART model only requires 4 variables. Therefore, in emergency situations where it takes time to collect these data, the hospital could use the CART model to quickly access if the patient has AHD.

The rationale behind why a model with a higher sensitivity and accuracy is that, we want to reduce the number of false negative. In the event of false negative, the patient is wrongly classified as non-AHD patient when in fact he is. A potential life might be lost to due to this misclassification especially if early detection is needed to cure. For a case of false positive the patient is classified as AHD patient but, he is not. After a few checks we can make sure that the patient does not have AHD. Therefore, the detrimental effects of misclassification are lesser for a model with a high false positive rate when compared to a model with a high false negative rate. The only downside of having a model with high false positive ratio is that, there might be additional cost incurred to check if a patient really has AHD. Therefore, a model which has both a high accuracy and high sensitivity needs to be selected, so that we can identify all potential AHD patients and at the same time reduce unnecessary cost.

**Part B**

Q5- Key finding to the Hospital

The logistic model obtained is shown in Figure 23. It can be observed that the factors that affect the outcome of AHD are ChestPain,RestBP,Chol,ExAng,Slope,Ca,gender and the newly created variable OldPeak\_per\_Chol. The statistically significant variables are Ca, Oldpeak\_per\_Chol,Slope,Thal and ChestPain. It also can be noted that not all the categories affect the outcome of AHD equally, when ChestPain is nonanginal and typical it is able to affect the outcome more than nontypical pain. Similar logic can be applied to gender. The estimate refers to how much the probability differs for one unit increase in the unit. An example is when Ca is increased by one unit, the log-odds of AHD will be increased by 1.6424. For categorical variables it will refer to how much the probability is different when its value is a 0 or 1. For example males (genderM = 1) has 1.0407 more in terms of log-odds[[2]](#footnote-2) of the AHD when compared to females (genderM = 0) .

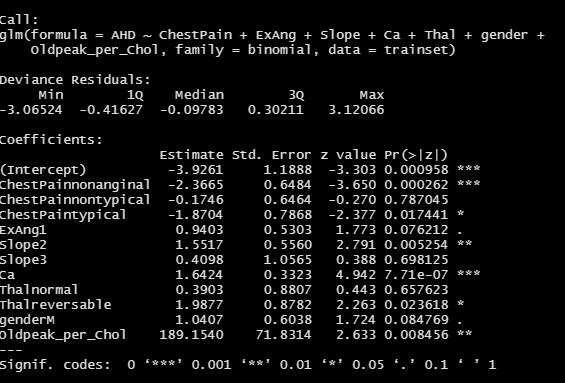


Figure 23: Summary of Logistic Regression Model

**Part C**

Q6- The missing information in CP table

The cross validation in question is measured in different methods depending on the type of the outcome variable. For a continuous outcome variable, the model is trained on 9 parts of the train set and used to predict the value of the outcome variable in the remaining part of the dataset. The difference between the actual and predicted is averaged out with the number of points in the test set will give us the error in one of the sets, this process is repeated 10 times. The average of all the error will give us the 10-fold cross validation error[[3]](#footnote-3). Therefore, we can quantify how much the predicted value deviates from the actual result.

For a categorical variable, the model is still trained on 9 parts of the train set and tested on the remaining part of the dataset. However, the error will now be treated as the number of misclassifications in one of the sets. Then the number of misclassified datapoints are average with the number of datapoints tested. This process is repeated 10 times and the average of these error will give us the 10-fold cross validation error3. Therefore, if we referred to the CP table obtained from training the CART model (Figure 24), at the size of 2 for the subtree, we can see that the xerror is 0.68041. This would mean that when the size of the subtree on average, the chances of misclassification is 0.68041. Therefore, if the dataset contains 100 points, 68 points will be misclassified.

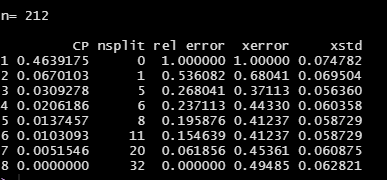
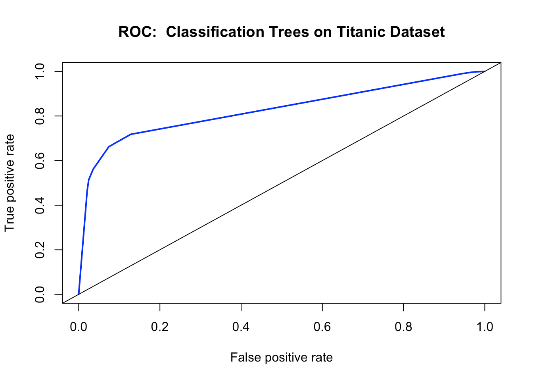


Figure 24: CP table of Q3 CART Model

But this misclassification can either be a false negative or a false positive. From the xerror we are only able to obtain the overall accuracy in the model. But we do not want to maximize overall accuracy in all situation. For example, in this problem we want to maximize sensitivity. Therefore, the information regarding the type of misclassification and its proportion is missing.

To obtain this information we can make use of a confusion matrix and calculate the average number of true positive, true negative, false positive and false negative across the 10 folds. Another way to see the performance of the model with respective to the true positive and true negative we can make use of the ROC library[[4]](#footnote-4). An example of the graph obtained using ROC is shown in figure 25. From the graph we can see how well the model is able to predict true positives and true negatives and its relationship.

Figure 25: Example of ROC curve obtained from logistic model

**Part C**

Q7- Comments on the Approach taken by the paper

The approach taken by the paper was to find the correlation between the variables in the dataset. Based on their correlation, the variables are given specific weight to them. Even though this is a suitable way to filter the specific variables to use in the modelling process, usually this approach is used more to pick out the independent numeric variables that affects the outcome variables.

If we looked at two numeric variables, we can quantify their relationship as either positive or negative correlation by increasing one of the variables and seeing the effect on the other variable. However, for categorical variables that do not hold numeric data, the correlation between these 2 variables will not hold any meaning.

But the categorical data in the dataset are mainly numeric in data type. Thus, we might be able to see the effect of the categorical variable on AHD. From the paper the most correlated variable with AHD is Thal. Therefore, when AHD increase from 0 to 1, Thal is expected to increase as well.

This approach is like how Netflix approached its rating. Even though Netflix’s rating is a categorical variable, they treated as if it was a numeric variable and calculated the RSME[[5]](#footnote-5) to evaluate the effectiveness of the model developed.

There can also be a few limitations to this approach. If we refer to the earlier example of the correlation between AHD and Thal. Even though we know there is a positive correlation, we will not know the level of Thal that affects AHD the most among the three levels.

Another limitation could be that correlation is not always mean causation. Correlation would just mean that there is a trend observed in the dataset, but these trends does not quantify that these variables case AHD.

In addition to ,that the model that was built on using the correlated variables with AHD might not be suitable for different datasets. This is because, the correlation between variables changes from sample to sample. Therefore, the features identified as correlated in this paper might not be correlated in another sample, resulting in inaccurate results.

In conclusion I believe that even though this approach might have some benefits to it ,there are too many loopholes present in the approach. They should conduct further testing on different datasets and even conduct a hypothesis testing on whether this approach is relevant and effective.

1. Adapted from <https://www.mayoclinic.org/healthy-lifestyle/fitness/in-depth/exercise-intensity/art-20046887#:~:text=You%20can%20calculate%20your%20maximum,beat%20per%20minute%20during%20exercise> on 8 November 2020 . [↑](#footnote-ref-1)
2. Adapted from <https://stats.idre.ucla.edu/stata/output/logistic-regression-analysis/> on 9 November 2020 [↑](#footnote-ref-2)
3. Adapted from <https://rpubs.com/malshe/212816> on 9 November 2020 [↑](#footnote-ref-3)
4. Adapted from <https://rstudio-pubs-static.s3.amazonaws.com/222569_a8d12e00f8204a479e84a33b49e54790.html> on 9 November 2020 [↑](#footnote-ref-4)
5. Adapted from BC2406 Exercise 2.1 solutions [↑](#footnote-ref-5)