Heart Disease Or Attack Project - Edx Choose Your Own Submission

Suhayl Hafiz

2024-05-21

Introduction

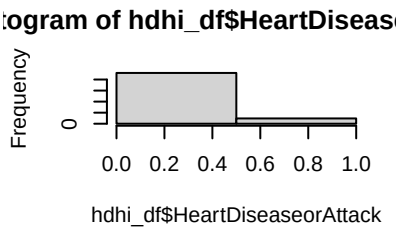
The following paragraph has been taken from the Kaggle website and has been reduced [1] (the link can be found in my reference section) to give you some context, Heart Disease is among the most prevalent chronic diseases in the United States which impacts millions of Americans each year and exerts significant financial burden on the economy. There are different types of coronary heart disease, the majority of individuals only learn they have the disease following symptoms such as chest pain, a heart attack, or sudden cardiac arrest. The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984.

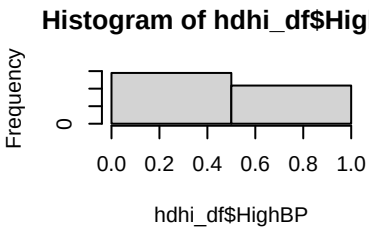
For this project, I've downloaded a csv dataset available on Kaggle for the year 2015 which has been cleaned. The dataset contains 253,680 survey responses and contains the following variables:

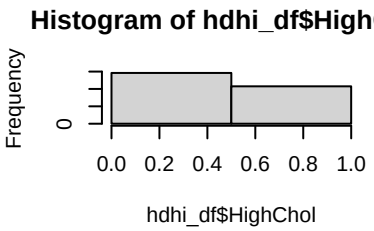
- HeartDiseaseorAttack: This is my target variable that I will be predicting Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI). This is binary variable which contains 23,893 people that have had heart disease or attack. Therefore we have an overall response rate of 9.419% i.e. we have an imbalance and will need to take this into account carrying out modelling.
- HighBP - High Blood Pressure:Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional. O means no and 1 means yes
- HighChol - High Cholesterol: Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high? O means no and 1 means yes.
- CholCheck: checked cholesterol in past 5 years by a doctor, nurse or other health professional that your blood cholesterol is high. O means no and 1 means yes.
- BMI - BMI: BMI to nearest percentage.
- Smoking - Smoker: Whether the person smoked atleast 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]. 0 means no and 1 means yes.
- Stroke: (Ever told) you had a stroke. O means no and 1 means yes.
- Diabetes - (Ever told) you have diabetes. 0 is for no diabetes or only during pregnancy, 1 is for pre-diabetes or borderline diabetes, 2 is for yes diabetes.
- PhysActivity - Physical Activity: Adults who reported doing physical activity or exercise during the past 30 days other than their regular job. 1 for physical activity, 0 for no physical activity.

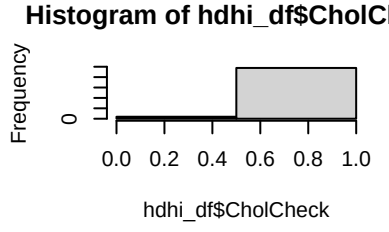
- Fruits: 0 means no fruit consumed per day. 1 means consumed 1 or more pieces of fruit per day.
- Veggies: 0.this means no vegetables consumed per day. 1 will mean consumed 1 or more pieces of vegetable per day.
- HvyAlcoholConsump - Alcohol Consumption: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week). 0 means not heavy drinker and 1 means heavy drinker.
- AnyHealthcare: Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service? . 0 means no and 1 means yes.
- NoDocbcCost: Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? . 0 means no and 1 means yes.
- GenHlth: Health General and Mental Health: - Would you say that in general your health is: –> GENHLTH. This is made of 5 categories, 1 is Excellent , 2 is Very Good, 3 is good, 4 is fair and 5 is poor.
- MentHlth: Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? in days from 0-30.
- PhysHlth: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? in days from 0-30.
- DiffWalk - Do you have serious difficulty walking or climbing stairs? –> DIFFWALK . 0 means no and 1 means yes.
- Sex: sex of respondent. female is 0 and male is 1.
- Age: age category - 1 is 18-24 all the way up to 13 wis 80 and older. 5 year increments.
- Education: What is the highest grade or year of school you completed? –> EDUCA. 1 being never attended school or kindergarten only up to 6 being college 4 years or more .
- Income: Is your annual household income from all sources: (If respondent refuses at any income level, code “Refused.”) –> INCOME2. 1 being less than $10,000 all the way up to 8 being $75,000 or more

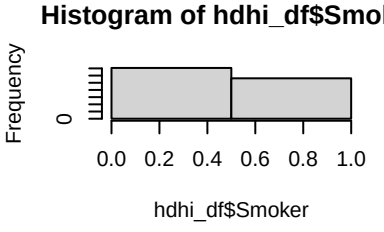
Here are histograms of all the variables listed above:

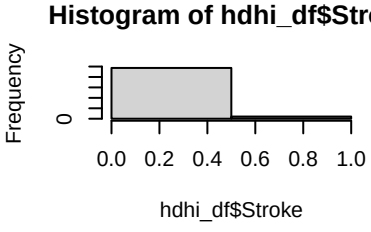

Frequency
0
0.0 0.2 0.4 0.6 0.8 1.0
hdhi_df$HeartDiseaseorAttack

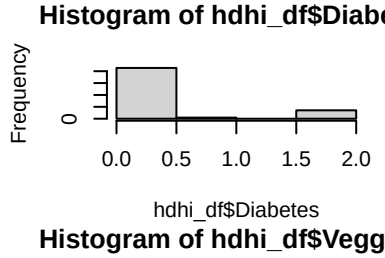

Histogram of hdhi_df$HighBP
Frequency
0
0.0 0.2 0.4 0.6 0.8 1.0
hdhi_df$HighBP

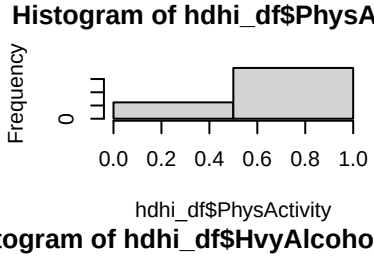

Histogram of hdhi_df$HighChol
Frequency
0
0.0 0.2 0.4 0.6 0.8 1.0
hdhi_df$HighChol

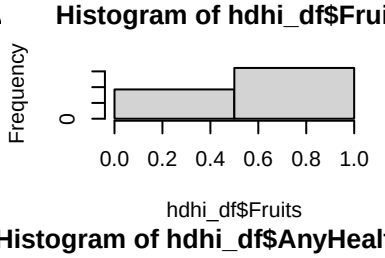

Histogram of hdhi_df$CholChe
Frequency
0
0.0 0.2 0.4 0.6 0.8 1.0
hdhi_df$CholCheck

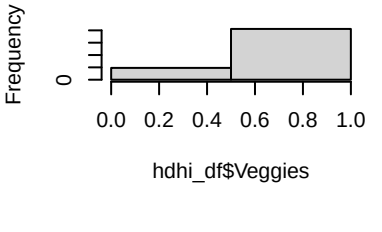

Histogram of hdhi_df$Smoke
Frequency
0
0.0 0.2 0.4 0.6 0.8 1.0
hdhi_df$Smoker

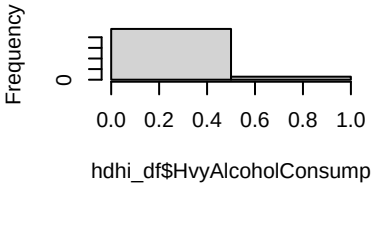

Histogram of hdhi_df$Stroke
Frequency
0
0.0 0.2 0.4 0.6 0.8 1.0
hdhi_df$Stroke

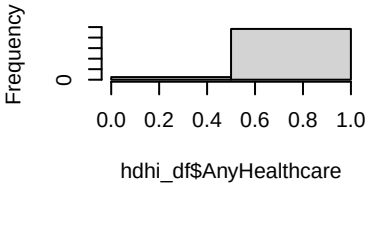

Histogram of hdhi_df$Diabete
Frequency
0
0.0 0.5 1.0 1.5 2.0
hdhi_df$Diabetes

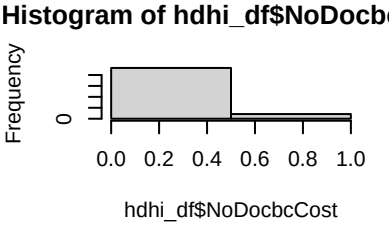

Histogram of hdhi_df$PhysActiv
Frequency
0
0.0 0.2 0.4 0.6 0.8 1.0
hdhi_df$PhysActivity

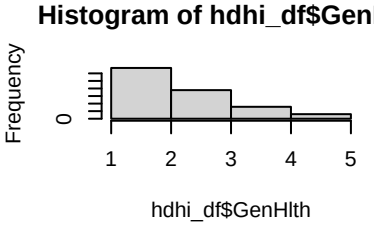

Histogram of hdhi_df$Fruits
Frequency
0
0.0 0.2 0.4 0.6 0.8 1.0
hdhi_df$Fruits

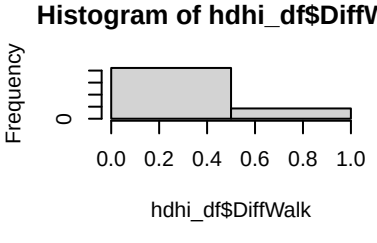

Histogram of hdhi_df$Veggies
Frequency
0
0.0 0.2 0.4 0.6 0.8 1.0
hdhi_df$Veggies

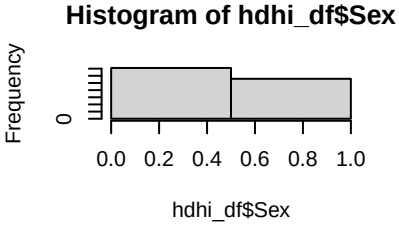

Histogram of hdhi_df$HvyAlcoholCo
Frequency
0
0.0 0.2 0.4 0.6 0.8 1.0
hdhi_df$HvyAlcoholConsump

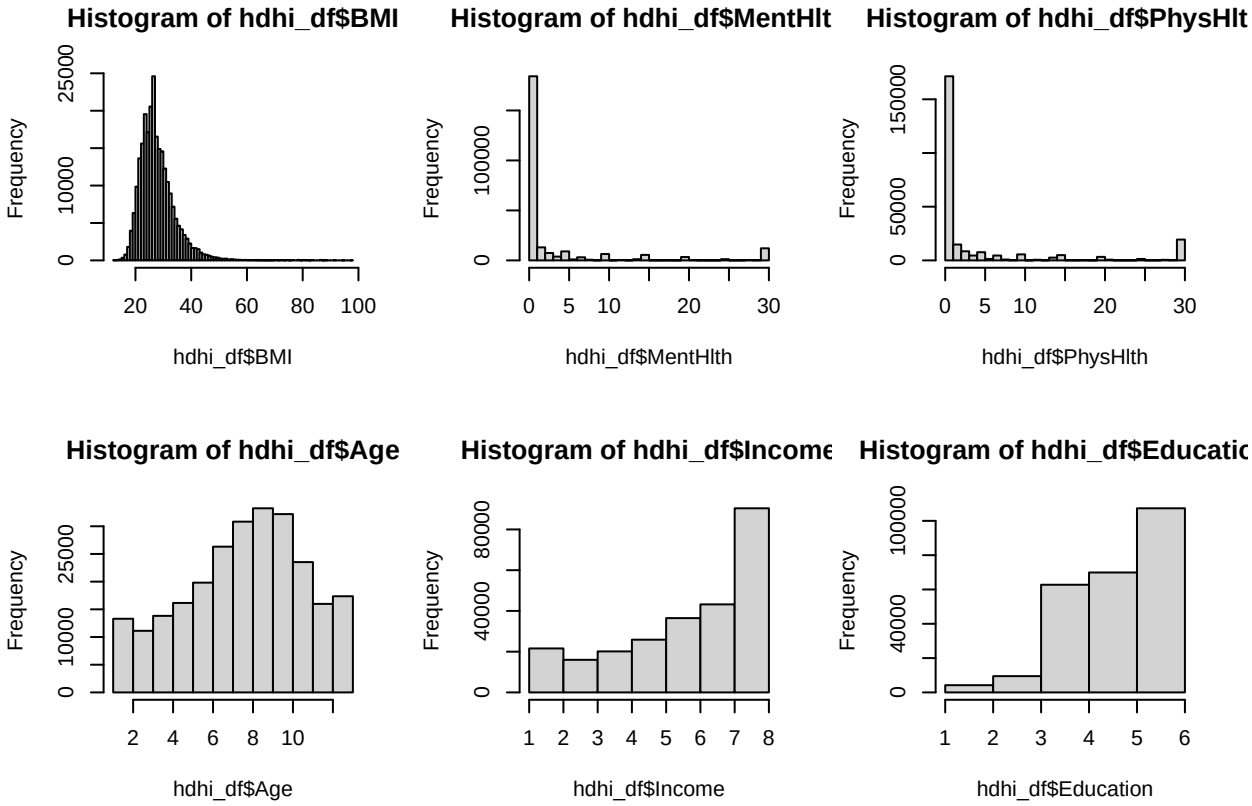

Histogram of hdhi_df$AnyHealth
Frequency
0
0.0 0.2 0.4 0.6 0.8 1.0
hdhi_df$AnyHealthcare


Histogram of hdhi_df$NoDocbcC
Frequency
0
0.0 0.2 0.4 0.6 0.8 1.0
hdhi_df$NoDocbcCost


Histogram of hdhi_df$GenHlth
Frequency
0
1 2 3 4 5
hdhi_df$GenHlth


Histogram of hdhi_df$DiffWalk
Frequency
0
0.0 0.2 0.4 0.6 0.8 1.0
hdhi_df$DiffWalk


Histogram of hdhi_df$Sex
Frequency
0
0.0 0.2 0.4 0.6 0.8 1.0
hdhi_df$Sex



The aim of this project is to create a predictive model that identifies which respondents have heart disease based and determines which people are most likely to have heart disease or attack based on the variables I described above.

Methods/Analysis

I decided to create 4 datasets which come from the original dataframe (hdhi_df) and each observation does not crossover into another dataset (except between Train & Train2).

1) Train - This represents 70% of the original dataframe and has been randomly selected. Going forward this dataset will be used for any exploratory analysis and to assess any models that have been built from my Train2 dataset.

2) Train2 - Since I have imbalance data with regards to Heart Disease or Attack, I can fix this by either adding in random samples of observations where heart disease is present or I can remove observations where heart disease isn't present. I have decided to under sample since this requires less computational requirements for model building. This dataset comes from my Train dataset only and it contains all observations where Heart Disease or Attack is present and I have randomly selected 20% of observations where Heart Disease or Attack is not present.

3) Test - This represents 15% of the original dataframe and has been randomly selected. Going forward this dataset will be used to compare any exploratory analysis and model building against my trainning set.

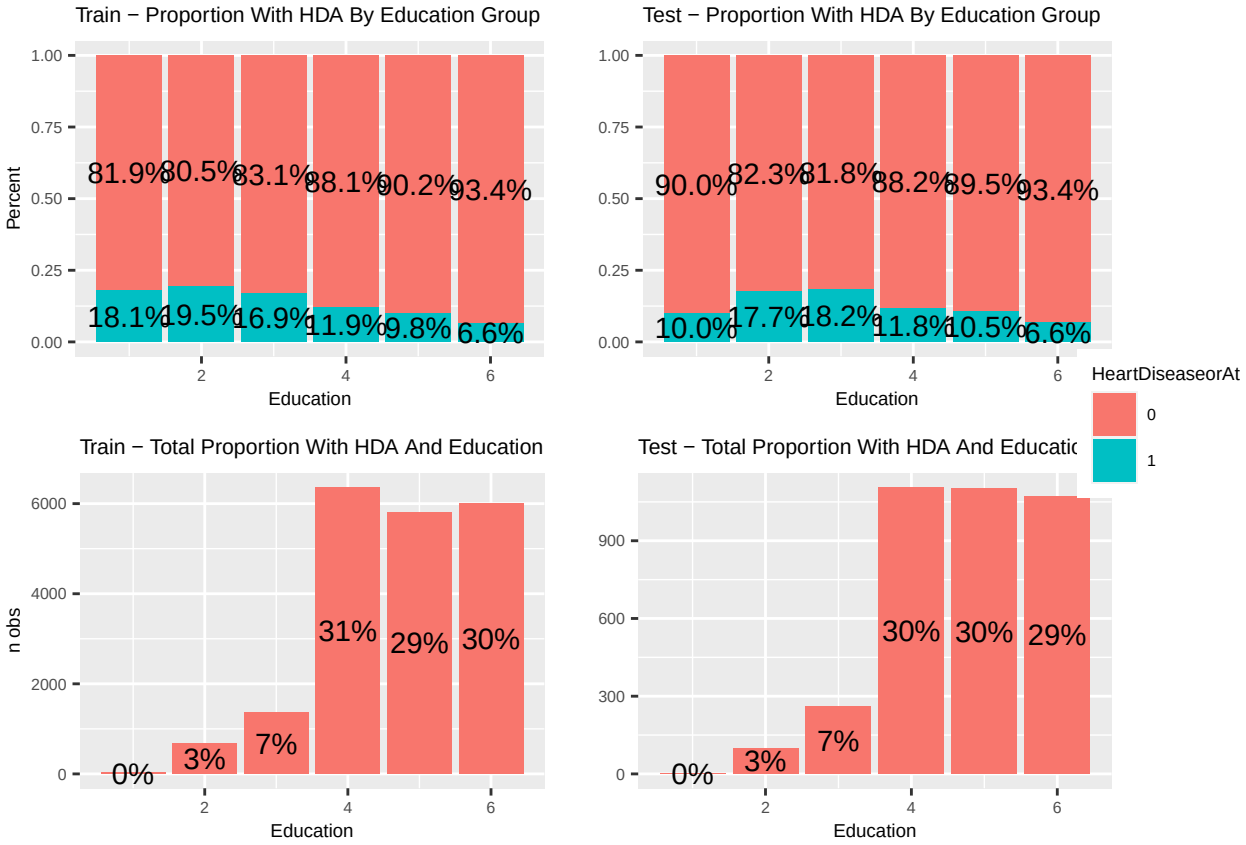
4) Holdout - This remaining 15% of the original dataframe goes intom my holdout dataset which will be used only at the final stage to determine how well my final model performed i.e. the records will not be used in any model building.

Here is a summary table of all the 4 datasets which shows you the number of records and the average response rate of observations with heart diesease or attack:

```
##   dataset N_Records    MeanHDA
## 1   Train    215628 0.09390246
## 2  Train2     59326 0.34130061
## 3    Test     38052 0.09578997
## 4 Holdout     38052 0.09352991
```

Exploratory Analysis

Before any model building was done I did exploratory analysis to first of all to understand my data more, to see if could create my own variables based on the variables I have and if I could simplify any variables by reducing the number of categories (all variables are categorical except BMI). I created a 2 by 2 graphical output of all variables to compare my Train and Test datasets which shows the average response rate of heart disease or attack by each value held in the variable that I'm exploring followed by the proportion of records held in each value by heart disease or attack.Here is an example below using the Education variable:



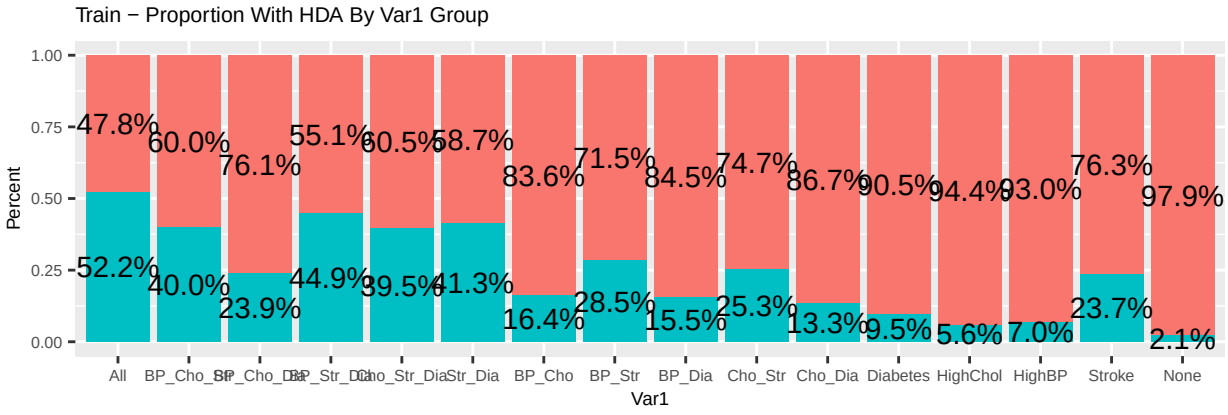
As you can see from the above chart I can reduce the number of categories in the Education variable if I group 1 and 3 levels together since their response rates to heart disease or attack are similar in the Train and Test datasets and kept groups 4, 5 & 6 separately i.e. I created a new variable called Education2 which will replace Education in the model building step.

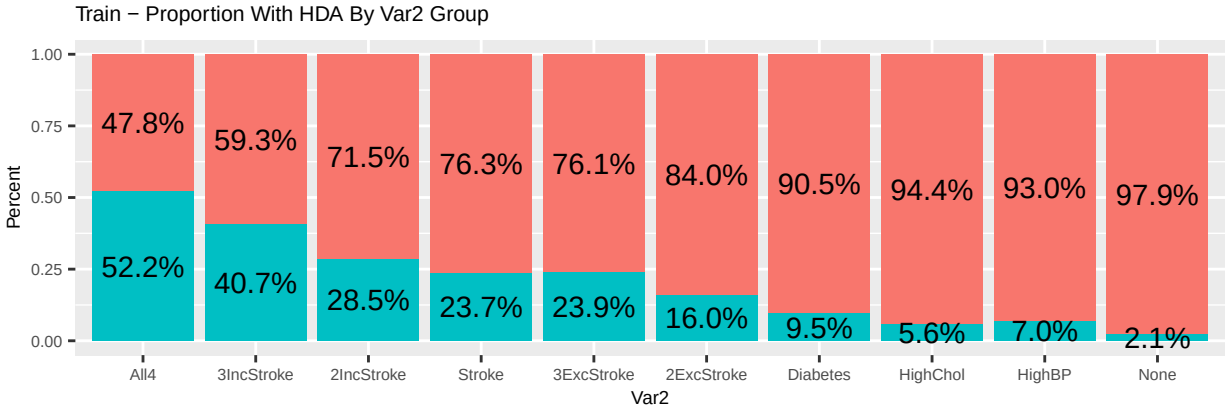
Based on the variables that I've been given (see introduction) I created these additional variables using the training and test datasets.

- Var1 - I have created this variable which is the outcome of the responses from HighBP, HighChol, Stroke and Diabetes variables.

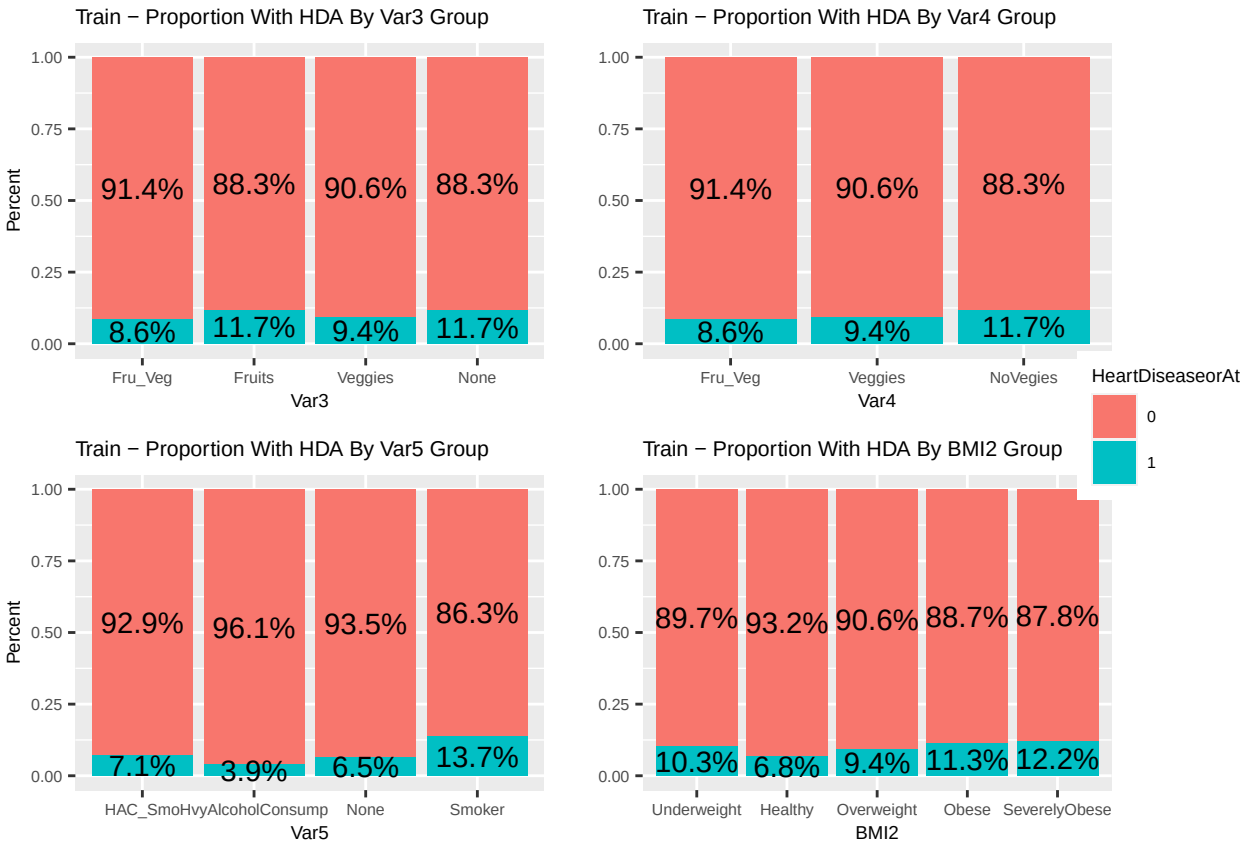
- Var2 - I have created this variable based on the information I gathered from Var1, which basically counts the no. of conditions a person has from HighBP, HighChol, Stroke and Diabetes but includes and excludes people with a stroke.
- Var3 - I have created this variable based on outcomes from Fruits and Veggies variables.This variable replaces the Fruits and Veggies variables when it came to the model building stage.
- Var4 - I have created this variable based on the information I gathered from Var3, which groups people who eat fruits and veggies consistently, have veggies and do not have veggies consistently. This variable replaces the Fruits and Veggies variables when it came to the model building stage.
- Var5 - I have created this variable which is the outcome of the responses from HvyAlcoholConsump, Smoker variables.
- BMI2- I have created this variable which is the outcome of the BMI variable, I have converted the numeric responses into bands which I got from the NHS [3]. Below 18.5 – underweight, 18.5 to 24.9 – healthy weight, 25 to 29.9 – overweight, 30 to 39.9 – obese and 40 or above – severely obese. This variable replaces the BMI variable when it came to the model building stage.
- MentHlth2 - I have created this variable which is based on the outcome MentHlth with the following categories 0, 1 to 9, 10 to 19 & 20 and over. This variable replaces the MentHlth variable when it came to the model building stage.
- PhysHlth2 - I have created this variable which is based on the outcome PhysHlth2 with the following categories 0 to 2, 3 to 9, 10 to 24 & 25 and over. This variable replaces the PhysHlth variable when it came to the model building stage.
- Education2 - I have created this variable which is based on the outcome from Education with the following categories 1 to 3, 4, 5 and 6. This variable replaces the Education variable when it came to the model building stage.
- Age2- I have created this variable which is based on the outcome from Age with the following categories 1 to 4, 5 to 6, 7, 8 to 9, 10, 11 to 12 and 13. This variable replaces the Age variable when it came to the model building stage.

I have provided you charts below to show you the response rates for by category for each of the categorical responses using my Train dataset only:





As you can see from the above charts when a person has had a stroke they more likely to suffer from heart disease or attack which you can see in the 1st chart above. Therefire it makes sense to create Var2 based on Var1 since the observations where a stoke is present they are more likely to suffer from heart disease or attack irregardless of how many conditions they have from diabetes, high cholesterol or high blood pressure.



Similarly with Var3 I saw that I could reduce the number of categories by creating Var4 since it seems that anyone who does not have veggies it more likely to suffer from heart disease or attack. With reagrds to Var5 I found it interesting that someone with high alcohol consumption is less likely to have heart disease or attack comnpare to someone who doesn't smoke or not had heart disease, this could just mean people who have high alcohol consumption will have liver issues before having heart disease or attack. Interstingly someone who is underwigeight is more likely to have heart disease or attack than someone who is classed as overweight.

With regards to mental health is seems as that having some mental health (between 1 and 9) has a positive effect on heart disease or attack. Not surprisingly we saw that people who exercise more frequently are less likely to have heart disease or attack. We saw that a person with lower education is more likely to have heart disease or attack however this could be linked to income and standard or living (people with high income are less likely). Not surprisingly the older you are the more likely you are to have heart disease or attack since the person has had more time to have heart disease or attack.

Model creation

In short form I decided to create 3 classification models using to different algorithm (Random Forest, GLM Logistic regression using forward selction based on AIC and CART) using the Train2 dataset. Each modek was assessed in the Train and Test dataset and based on the best model create out of the 3 (GLM) I then proceeded to see if I could improve the model further by creating dummy variables (converting a single cateogorical variable into multiple variables with a binary response) of the variables selected in my GLM model and again running the forward selection as my 2nd model. Then I created a 3rd GLM model that uses backward selecion based on my 2nd model. I then created a 4th GLM model that uses forward and backwards selection of the variables used in my 2nd model. A 5th GLM model created by removing any insignificant variables (so that all variables incuded in the model have a p-value of <0.05) from the 3rd model and then a 6th GLM model was created which is the same as the 4th GLM model but all insignifcant variables. I then compared the results from the GLM models based on the Train and Test datasets I concluded that that the 5th GLM model is the best since there was minimal differences between the results (between accuracy, sensitivity, precision and Somers D) but it had the lowest number of variables. I then proceed to adjuist my cuttoff value in the 5th GLM model that determines to identify people with heart disease or attack, the best cuttoff value with determined by the highest Somers D statistic. I then assessed my optimised 5th GLM model against my holdout sample which produced similar results compared to the Train and Test datasets.

The metrics I used to determine how each model perfomed on the Train and Test datasets:

- NVars - This is the number of variables used in the model, the lower the better.
- Accuracy - Shows the proportion of correctly classified outcomes, the higher the better but because I have an imbalanced dataset with regards to heart disease or attack it's acceptable to reduce this figure if I'm identifying more people with heart disease or attack.
- Sensitivity - This is the true positives divided by the true positives plus the false negatives, the higher the better.
- Precision - This is the true positives divided by the true positives plus the false positives, the higher the better.
- NcountTP - The number of true positives identified in the model i.e. the higher the better.
- Somers D - This is a measure of the strength and direction of the association between an ordinal dependent variable and an ordinal independent variable [4]. the score ranges between -1 and 1 i.e. the higher the better and holds the most weight with regards to picking the best model.

Results

Here is the table of my summary results based on the 3 initial models I created (Random Forest, GLM & CART) based on the train and test datasets:

```
##   Model NVars TrainAccuracy TestAccuracy TrainSensitivity TestSensitivity
## 1    RF    75     0.8565910    0.8520709        0.5241505       0.5075446
## 2   GLM    46     0.8369646    0.8352518        0.6140853       0.6197531
## 3  CART    20     0.8102380    0.8096026        0.5267681       0.5412894
##   TrainPrecision TestPrecision TrainNcountTP TestNcountTP TrainSomersD
## 1      0.3326855      0.325475         10613         1850    0.4151936
## 2      0.3126084      0.316298         12434         2259    0.4741478
## 3      0.2539403      0.261463         10666         1973    0.3663832
##   TestSomersD
## 1   0.3961138
## 2   0.4778343
## 3   0.3793166
```

From the above table you will see that the GLM model performed the best when it came to the Somers D statistic plus it also identified the most true positives i.e. I explored GLM logistic regression further.

As stated in the first paragraph of this section I created another 3 GLM models based on the variables from my inintal model except that I had used various selection methods (forward, backwards and both directions). I found that there was no change to the model between the forward and backward models. I could see that there are insignificant variables i.e. I created another 2 GLM models that removed the insignificant variables.

Here is the table of my summary results based on the forward (GLM2), backward (GLM3), both direction (GLM4), removed insigniifcant variables from GLM3 (GLM5) and removed insiginficant variables from GLM4 (GLM6) models based on the train and test datasets:

```
##   Model NVars TrainAccuracy TestAccuracy TrainSensitivity TestSensitivity
## 4  GLM2    41     0.8369321    0.8347262        0.6151225       0.6197531
## 5  GLM3    41     0.8369321    0.8347262        0.6151225       0.6197531
## 6  GLM4    51     0.8369089    0.8346210        0.6150731       0.6203018
## 7  GLM5    36     0.8373217    0.8352780        0.6137396       0.6200274
## 8  GLM6    45     0.8370620    0.8354620        0.6134433       0.6222222
##   TrainPrecision TestPrecision TrainNcountTP TestNcountTP TrainSomersD
```

```
## 4      0.3127511     0.3154147         12455         2259    0.4750416
## 5      0.3127511     0.3154147         12455         2259    0.4750416
## 6      0.3127024     0.3153417         12454         2261    0.4749718
## 7      0.3131489     0.3163937         12427         2260    0.4742320
## 8      0.3126510     0.3171141         12421         2268    0.4736798
##   TestSomersD
## 4   0.4772530
## 5   0.4772530
## 6   0.4776273
## 7   0.4781086
## 8   0.4802744
```

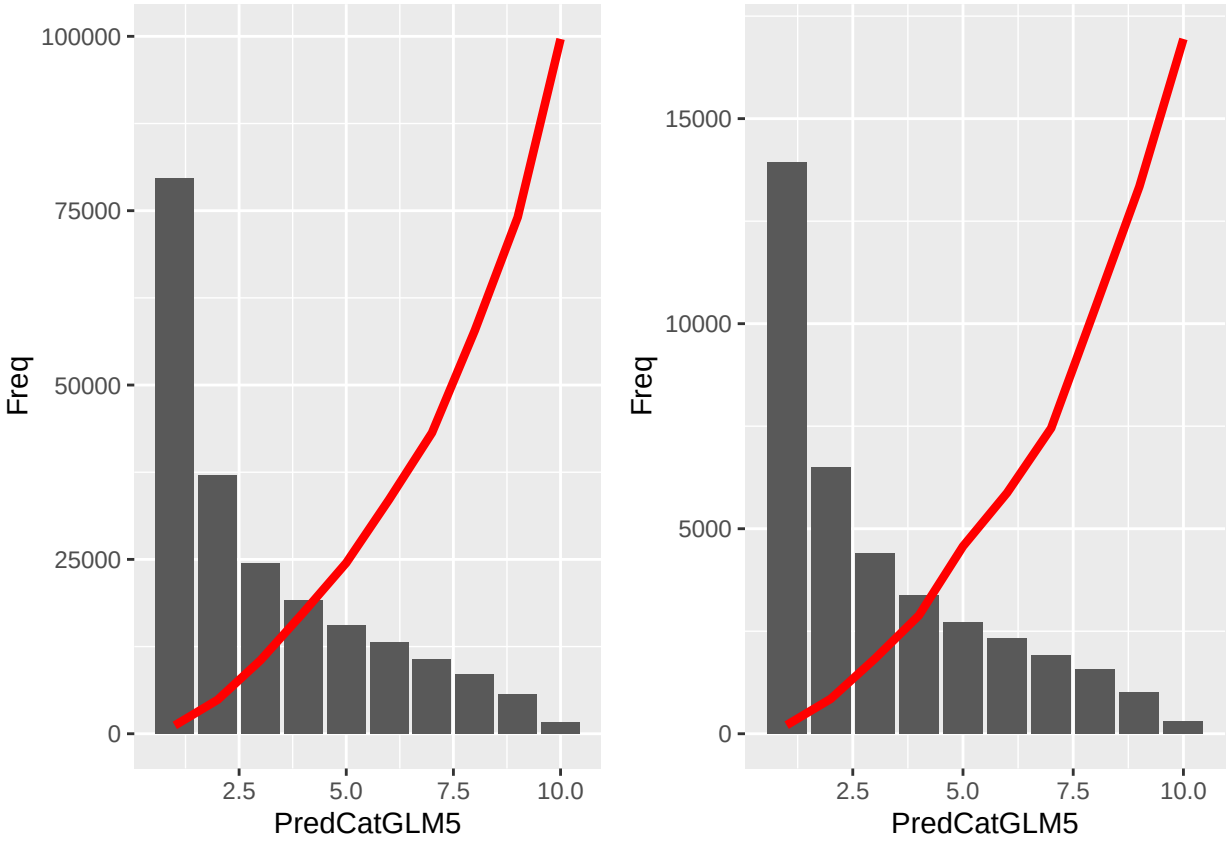
From the above table you will see that the results are very similar to each except that the GLM5 model used fewer (removed insignificant variables from GLM3) number of variables i.e. I believe this is the best model I have and have decided this to be my final model. Look below to the see the model summary of my GLM5:

```
##
## Call:
## NULL
##
## Coefficients:
##                   Estimate Std. Error z value             Pr(>|z|)
## (Intercept)      -1.38789    0.11659 -11.904 < 0.0000000000000002 ***
## Var1_None        -0.48162    0.06845  -7.036  0.00000000000197527 ***
## DiffWalk          0.30166    0.02793  10.800 < 0.0000000000000002 ***
## Sex               0.80409    0.02239  35.908 < 0.0000000000000002 ***
## Age2_1_4         -1.43018    0.06334 -22.581 < 0.0000000000000002 ***
## Stroke            0.73574    0.05424  13.565 < 0.0000000000000002 ***
## Age2_5_6         -0.80090    0.04526 -17.696 < 0.0000000000000002 ***
## GenHlth_1        -1.82538    0.06034 -30.253 < 0.0000000000000002 ***
## GenHlth_2        -1.41774    0.04909 -28.881 < 0.0000000000000002 ***
## GenHlth_3        -0.91355    0.04496 -20.321 < 0.0000000000000002 ***
## Age2_13           1.19966    0.03781  31.728 < 0.0000000000000002 ***
## Age2_11_12        0.78191    0.03016  25.926 < 0.0000000000000002 ***
## Var5_Smoker       0.37788    0.02162  17.478 < 0.0000000000000002 ***
## HighChol          0.42482    0.04288   9.907 < 0.0000000000000002 ***
## Var1_HighChol    -0.30954    0.05904  -5.243  0.00000015824543846 ***
## Age2_10           0.42967    0.03302  13.012 < 0.0000000000000002 ***
## GenHlth_4        -0.37383    0.04361  -8.573 < 0.0000000000000002 ***
## Age2_7           -0.36642    0.04295  -8.532 < 0.0000000000000002 ***
## Var1_HighBP      -0.33858    0.05433  -6.232  0.00000000046038339 ***
## Income_8         -0.12685    0.02766  -4.586  0.00000452684921435 ***
## HighBP            0.37522    0.04775   7.859  0.00000000000000388 ***
## NoDocbcCost       0.20900    0.03794   5.509  0.00000003605805993 ***
## CholCheck         0.48049    0.08366   5.744  0.00000000927039994 ***
## PhysHlth2_10to24  0.12876    0.03887   3.313             0.000924 ***
## Diabetes_2        0.11506    0.03290   3.498             0.000469 ***
## Income_2          0.21520    0.04445   4.842  0.00000128742013071 ***
## Var1_Stroke       0.57121    0.13366   4.274  0.00001924175576670 ***
## Income_1          0.22566    0.05065   4.456  0.00000836235360329 ***
## PhysHlth2_0to2   -0.09316    0.02865  -3.252             0.001147 **
## Education2_5      0.07755    0.02373   3.269             0.001081 **
## Income_3          0.14722    0.04028   3.655             0.000257 ***
## Var1_Str_Dia      0.82392    0.26178   3.147             0.001648 **
```

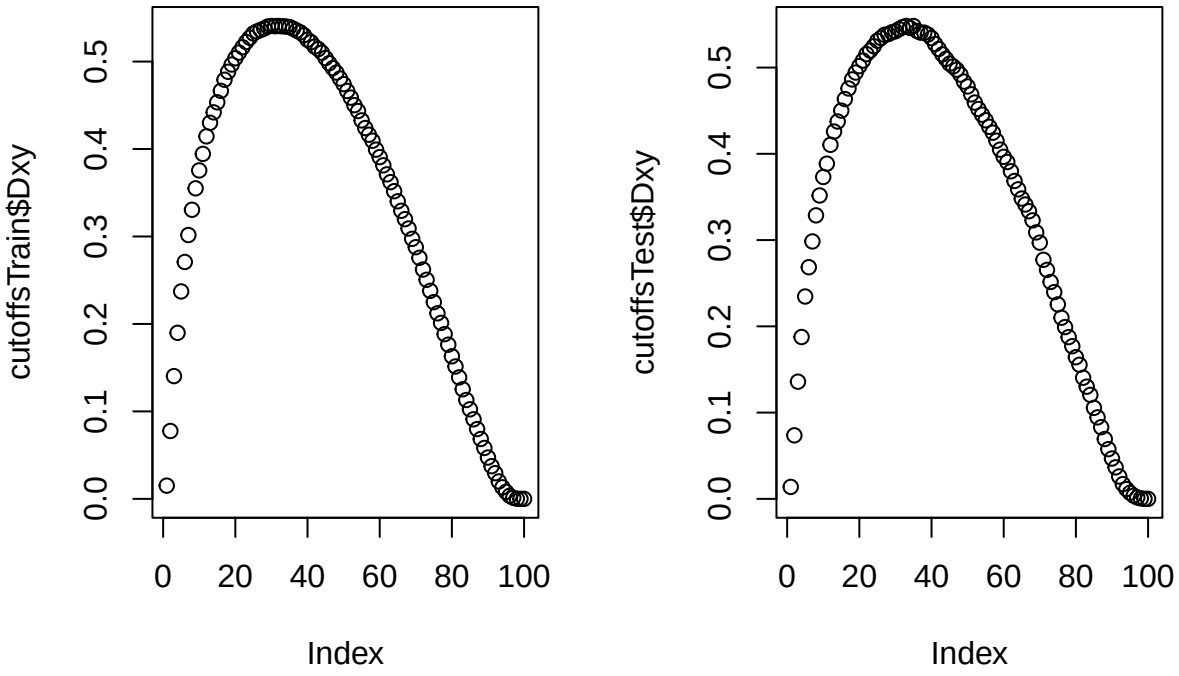
```
## Var1_BP_Str_Dia   0.47364    0.16653   2.844            0.004452 **
## BMI2_Healthy     -0.06911    0.02705  -2.555            0.010624 *
## Income_4          0.08321    0.03725   2.234            0.025496 *
## Var3_Fruits      -0.08452    0.03856  -2.192            0.028380 *
## Var1_All          0.21188    0.10231   2.071            0.038364 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 76162  on 59325  degrees of freedom
## Residual deviance: 54163  on 59289  degrees of freedom
## AIC: 54237
##
## Number of Fisher Scoring iterations: 5
```

Here is a graphical view (Train - LHS and Test - RHS) below of the performance of the Train and Test data, essentially I have scored all records in the dataset, I then put the probability scores into 10 bins (1 to 10) and then added the proportion of people that have been identified as having heart disease or attack. You will see the higher the probability score is the better the model perfomed.

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



I proceeded to fine tune my cuttoff value (what the probability score needs to be) that determines whether a person should be classified as having heart disease or attack. To do this I ran a sequence between 0.01 to 1 in increments of 0.01 which I effectively changed the cuttoff value and then I produced a list of statistics. I based my cuttoff value on the highest Somers D statistic on the Train & Test dataset where I took the average best cuttoff values. Here is a graphical view.




```
##            C       Dxy      n Missing
## 30 0.7703609 0.5407218 215628       0
```

```
##            C       Dxy     n Missing
## 33 0.7741212 0.5482424 38052       0
```

I then updated my model summary table and will see that my optimised 5th GLM model has perfomed best with regards to Somers D and identifying the most true positives.

```
##            Model NVars TrainAccuracy TestAccuracy TrainSensitivity
## 1             RF    75     0.8565910    0.8520709        0.5241505
## 2            GLM    46     0.8369646    0.8352518        0.6140853
## 3           CART    20     0.8102380    0.8096026        0.5267681
## 4           GLM2    41     0.8369321    0.8347262        0.6151225
## 5           GLM3    41     0.8369321    0.8347262        0.6151225
## 6           GLM4    51     0.8369089    0.8346210        0.6150731
## 7           GLM5    36     0.8373217    0.8352780        0.6137396
## 8           GLM6    45     0.8370620    0.8354620        0.6134433
## 9 GLM5 Optimised    36     0.7313290    0.7308157        0.8183524
##   TestSensitivity TrainPrecision TestPrecision TrainNcountTP TestNcountTP
## 1       0.5075446      0.3326855     0.3254750         10613         1850
## 2       0.6197531      0.3126084     0.3162980         12434         2259
## 3       0.5412894      0.2539403     0.2614630         10666         1973
## 4       0.6197531      0.3127511     0.3154147         12455         2259
## 5       0.6197531      0.3127511     0.3154147         12455         2259
## 6       0.6203018      0.3127024     0.3153417         12454         2261
```

```
## 7       0.6200274      0.3131489     0.3163937         12427         2260
## 8       0.6222222      0.3126510     0.3171141         12421         2268
## 9       0.8235940      0.2339569     0.2382162         16570         3002
##   TrainSomersD TestSomersD
## 1    0.4151936   0.3961138
## 2    0.4741478   0.4778343
## 3    0.3663832   0.3793166
## 4    0.4750416   0.4772530
## 5    0.4750416   0.4772530
## 6    0.4749718   0.4776273
## 7    0.4742320   0.4781086
## 8    0.4736798   0.4802744
## 9    0.5406628   0.5445810
```

I then ran my 5th GLM model using my cuttoff value against my holdout dataset and got similar results compared to my Train and Test data.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 25016   634
##          1  9477  2925
##
##                Accuracy : 0.7343
##                  95% CI : (0.7298, 0.7387)
##     No Information Rate : 0.9065
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2588
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##             Sensitivity : 0.7252
##             Specificity : 0.8219
##          Pos Pred Value : 0.9753
##          Neg Pred Value : 0.2358
##              Prevalence : 0.9065
##          Detection Rate : 0.6574
##    Detection Prevalence : 0.6741
##       Balanced Accuracy : 0.7736
##
##        'Positive' Class : 0
##

##             C           Dxy             n       Missing
##     0.7735543     0.5471087 38052.0000000     0.0000000
```

Conclusion

My final model has a moderate Somers D statistic value and was able to capture the majority of people who have heart disease or attack. At the same time I also predicted many people that would have it but haven't

had it however I have identified these people most at risk of having heart disease or attack and suggest that these people are examined by a medical professional to determine the risk they are at of having heart disease or attack.

If I were to do the project again I would look to create multiple models for based on the age group a person sits in as we know that the older a person is the more likely they are to have heart disease or attack and the at the same time nobody has control over this factor.

References

1) I got my dataset from kaggle from the following URL: https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset

2) To download the Kaggle dataset as a csv file, use the link below: https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset/download?datasetVersionNumber=3

3) BMI categories coming from the NHS UK which have been used to create the BMI2 variable: https://www.nhs.uk/conditions/obesity/

4) What is Somers' D:

   https://www.statology.org/somers-d/