Final Report for BIOS 611

Ashkan Habib

2023-12-14

Final Report

This report is based on the analysis of “Cardiovascular Diseases Risk Prediction Dataset The 2021 BRFSS Dataset from CDC.You can download the dataset for free and check it out yourself from keggle. The link is in the repository ReadMe file.

Overview

I was interested in this dataset because I generally like to conduct clinical research, especially on cardiovascular diseases. My question was whether I could predict the risk of cardiovascular health using other variables. Let’s look at the summary statistics of the study:

```r
load("~/work/tables/first_look.RData")
look
```

```
##  General_Health       Checkup            Exercise         Heart_Disease
##  Length:308854      Length:308854      Length:308854      Length:308854
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Skin_Cancer        Other_Cancer        Depression          Diabetes
##  Length:308854      Length:308854      Length:308854      Length:308854
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   Arthritis             Sex            Age_Category        Height_(cm)
##  Length:308854      Length:308854      Length:308854      Min.   : 91.0
##  Class :character   Class :character   Class :character   1st Qu.:163.0
##  Mode  :character   Mode  :character   Mode  :character   Median :170.0
##                                                           Mean   :170.6
##                                                           3rd Qu.:178.0
##                                                           Max.   :241.0
##   Weight_(kg)          BMI        Smoking_History    Alcohol_Consumption
##  Min.   : 24.95   Min.   :12.02   Length:308854      Min.   : 0.000
##  1st Qu.: 68.04   1st Qu.:24.21   Class :character   1st Qu.: 0.000
##  Median : 81.65   Median :27.44   Mode  :character   Median : 1.000
##  Mean   : 83.59   Mean   :28.63                      Mean   : 5.096
##  3rd Qu.: 95.25   3rd Qu.:31.85                      3rd Qu.: 6.000
##  Max.   :293.02   Max.   :99.33                      Max.   :30.000
##  Fruit_Consumption Green_Vegetables_Consumption FriedPotato_Consumption
##  Min.   :  0.00    Min.   :  0.00               Min.   :  0.000
##  1st Qu.: 12.00    1st Qu.:  4.00               1st Qu.:  2.000
##  Median : 30.00    Median : 12.00               Median :  4.000
##  Mean   : 29.84    Mean   : 15.11               Mean   :  6.297
##  3rd Qu.: 30.00    3rd Qu.: 20.00               3rd Qu.:  8.000
##  Max.   :120.00    Max.   :128.00               Max.   :128.000
```

```
load("~/work/tables/proportions.RData")
proportions
```

```
## $General_Health
##
##              freq    perc
##
## Excellent  55’954   18.1%
## Fair       35’810   11.6%
## Good       95’364   30.9%
## Poor       11’331    3.7%
## Very Good 110’395   35.7%
## NA              0    0.0%
##
## $Checkup
##
##                            freq    perc
##
## 5 or more years ago      13’421    4.3%
## Never                     1’407    0.5%
## Within the past 2 years  37’213   12.0%
## Within the past 5 years  17’442    5.6%
## Within the past year    239’371   77.5%
## NA                           0    0.0%
##
## $Exercise
##
##        freq    perc
##
## No   69’473   22.5%
## Yes 239’381   77.5%
## NA        0    0.0%
##
## $Heart_Disease
##
##        freq    perc
##
## No  283’883   91.9%
## Yes  24’971    8.1%
## NA        0    0.0%
##
## $Skin_Cancer
##
##        freq    perc
##
## No  278’860   90.3%
## Yes  29’994    9.7%
## NA        0    0.0%
##
## $Other_Cancer
##
##        freq    perc
##
## No  278’976   90.3%
## Yes  29’878    9.7%
## NA        0    0.0%
##
## $Depression
##
##        freq    perc
##
## No  246’953   80.0%
## Yes  61’901   20.0%
## NA        0    0.0%
##
## $Diabetes
##
##        freq    perc
##
## No  266’037   86.1%
## Yes  42’817   13.9%
## NA        0    0.0%
##
## $Arthritis
##
##        freq    perc
##
## No  207’783   67.3%
## Yes 101’071   32.7%
## NA        0    0.0%
##
## $Sex
##
##           freq    perc
##
## Female 160’196   51.9%
## Male   148’658   48.1%
## NA           0    0.0%
##
## $Age_Category
##
##         freq   perc
##
## 18-24 18’681   6.0%
## 25-29 15’494   5.0%
## 30-34 18’428   6.0%
## 35-39 20’606   6.7%
## 40-44 21’595   7.0%
## 45-49 20’968   6.8%
## 50-54 25’097   8.1%
## 55-59 28’054   9.1%
## 60-64 32’418  10.5%
## 65-69 33’434  10.8%
## 70-74 31’103  10.1%
## 75-79 20’705   6.7%
## 80+   22’271   7.2%
## NA         0   0.0%
##
## $Smoking_History
##
##        freq    perc
##
## No  183’590   59.4%
## Yes 125’264   40.6%
## NA        0    0.0%
##
## $‘Heart Disease‘
##
##        freq    perc
##
## No  283’883   91.9%
## Yes  24’971    8.1%
## NA        0    0.0%
##
## $‘General Health‘
##
##                   freq    perc
##
## Fair or worse   47’141   15.3%
## Good or better 261’713   84.7%
## NA                   0    0.0%
##
## $‘Check up‘
##
##                                   freq    perc
##
## Never or more than 5 years ago  14’828    4.8%
## Within the past 5 years        294’026   95.2%
## NA                                   0    0.0%
##
## $‘Female Sex‘
##
##        freq    perc
##
## No  148’658   48.1%
## Yes 160’196   51.9%
## NA        0    0.0%
##
## $Age
##
##                       freq    perc
##
## Less than 55 years 140’869   45.6%
## More than 55 years 167’985   54.4%
## NA                       0    0.0%
```

```
##
## $Smoking
##
##          freq    perc
##
## No   183'590   59.4%
## Yes  125'264   40.6%
## NA         0    0.0%
```

There are 308854 observations in this study. Mean BMI is 28.6, with 25.7% and and 18.1% having very good or excellent health status'. The study is about half female and male and 8.1% have heart disease.

Risk Factors

Since this a cross sectional look into the association of heart disease with its risk factors, let's look at the prevalence of each risk factor by their heart health status:

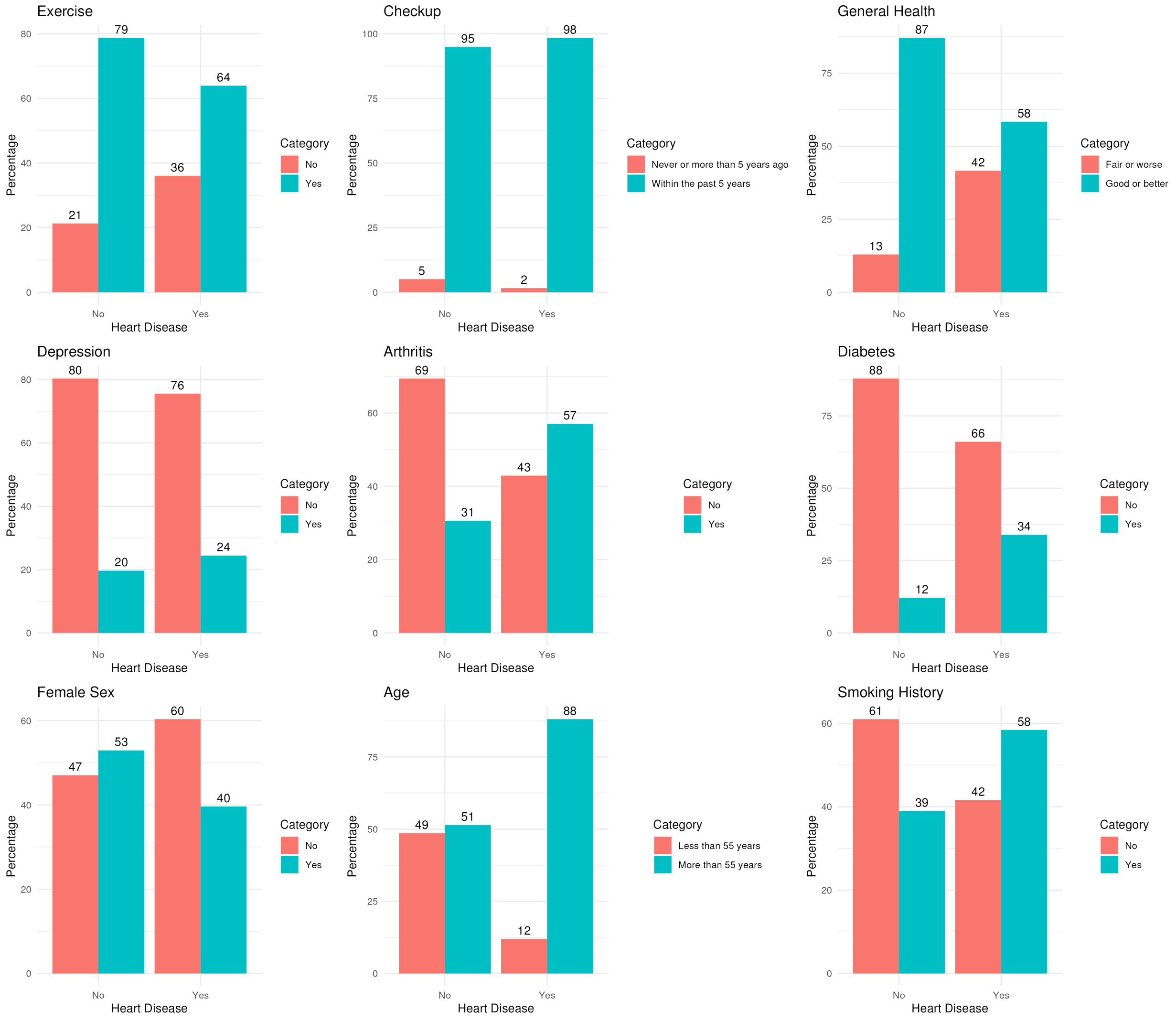

Figure 1: Prevalence of risk factors by heart status

Right at the start we notice that there are massive differences in arthritis, smoking and age. Difference in diabetes is also noticeable. Let’s look at the mean and median of continuous factors by heart status:

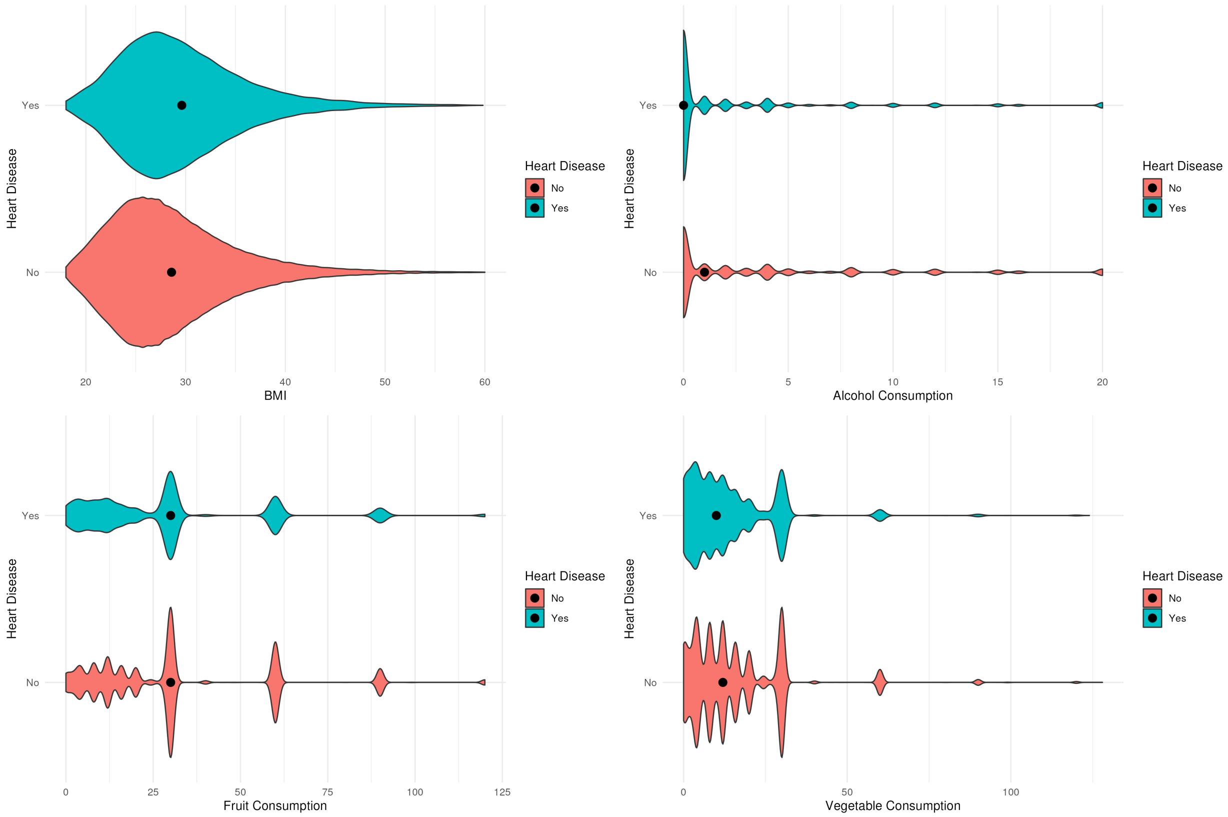

Figure 2: Distribution of of continuous risk factors by heart status

Here, we see thatalcohol consumption distribution clearly differs based on heart status, followed by a much smaller difference in BMI. The rest don’t look that different. Let’s look at their differnece individually and conduct a simple wilcoxon ramk sum test.

```
load("~/work/tables/BMI.RData")
BMI
```

```
## # A tibble: 2 x 3
##   Heart_Disease  Mean    SD
##   <chr>         <dbl> <dbl>
## 1 No             28.5  6.51
## 2 Yes            29.6  6.58
```

```
BMI2
```

```
##
##  Two Sample t-test
##
## data:  BMI by Heart_Disease
## t = -23.733, df = 308852, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -1.1051324 -0.9365252
## sample estimates:
##  mean in group No mean in group Yes
##          28.54368          29.56450
```

Based on above, there's a very small diffrence in BMI and it is statistically significant.

```
load("~/work/tables/alcohol.RData")
alcohol
```

```
## # A tibble: 2 x 3
##   Heart_Disease Median   IQR
##   <chr>          <dbl> <dbl>
## 1 No                 1     7
## 2 Yes                0     4
```

```
alcohol2
```

```
## 
##  Wilcoxon rank sum test with continuity correction
## 
## data:  Alcohol_Consumption by Heart_Disease
## W = 4082698008, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Alcohol is also different, with those having a heart disease consuming less alcohol. This is also statistically significant.

```
load("~/work/tables/fruit.RData")
fruit
```

```
## # A tibble: 2 x 3
##   Heart_Disease Median   IQR
##   <chr>          <dbl> <dbl>
## 1 No                30    18
## 2 Yes               30    20
```

```
fruit2
```

```
## 
##  Wilcoxon rank sum test with continuity correction
## 
## data:  Fruit_Consumption by Heart_Disease
## W = 3714264712, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Fruit looks to be similar but the p value is statistically significant, meaning there's still some difference however small.

```
load("~/work/tables/veggies.RData")
veggies
```

```
## # A tibble: 2 x 3
##   Heart_Disease Median   IQR
##   <chr>          <dbl> <dbl>
## 1 No                12    16
## 2 Yes               10    16
```

```
veggies2
```

```
## 
##  Wilcoxon rank sum test with continuity correction
## 
## data:  Green_Vegetables_Consumption by Heart_Disease
## W = 3748407442, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Vegetable consumption also looks to be similar but the p value is statistically significant.

```
load("~/work/tables/fries.RData")
fries
```

```
## # A tibble: 2 x 3
##   Heart_Disease Median   IQR
##   <chr>          <dbl> <dbl>
## 1 No                 4     6
## 2 Yes                4     7
```

```
fries2
```

```
## 
##  Wilcoxon rank sum test with continuity correction
## 
## data:  FriedPotato_Consumption by Heart_Disease
## W = 3720298992, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Same results are seen for fried potato consumption: very similar but with statistically significant p value.

Modeling

The next step is to perform a gradient boosting model to predict heart disease with all of the revelant risk factors inserted. For this step, I included 80% of the observations that were randomly chosen to fit (train) the model, we will compare them against the remaining 20% as a test. I also did not put cancer as a variable as it might lead to selection bias. I also did not put weight and weight because their data is alrerady captured in the BMI variable.

It's very appearant that age, general health and diabetes have high influence. This was very expected as all are massive risk factors for heart disease. But let's look deeper and see what variables had little to no influence:

```
load("~/work/tables/model_summary.RData")
model_sum
```

```
##                                                     var     rel.inf
## AgeCategory                                 AgeCategory 39.18660080
## GeneralHealth                             GeneralHealth 37.75661080
## Diabetes                                       Diabetes 11.21612661
## Sex                                                 Sex  4.86051572
## SmokingHistory                           SmokingHistory  3.15208224
## Arthritis                                     Arthritis  3.09821494
## Checkup                                         Checkup  0.43252879
## AlcoholConsumption                   AlcoholConsumption  0.20153094
## Depression                                   Depression  0.09578916
## Exercise                                       Exercise  0.00000000
## BMI                                                 BMI  0.00000000
## FruitConsumption                       FruitConsumption  0.00000000
## GreenVegetablesConsumption GreenVegetablesConsumption  0.00000000
## FriedPotatoConsumption           FriedPotatoConsumption  0.00000000
```

Interestingly, a lot of risk factors of heart disease have no influence in the model, including exercise, BMI and fried potato consumption. I suspect that the reason is because their influence is already captured by the "General Health" status.

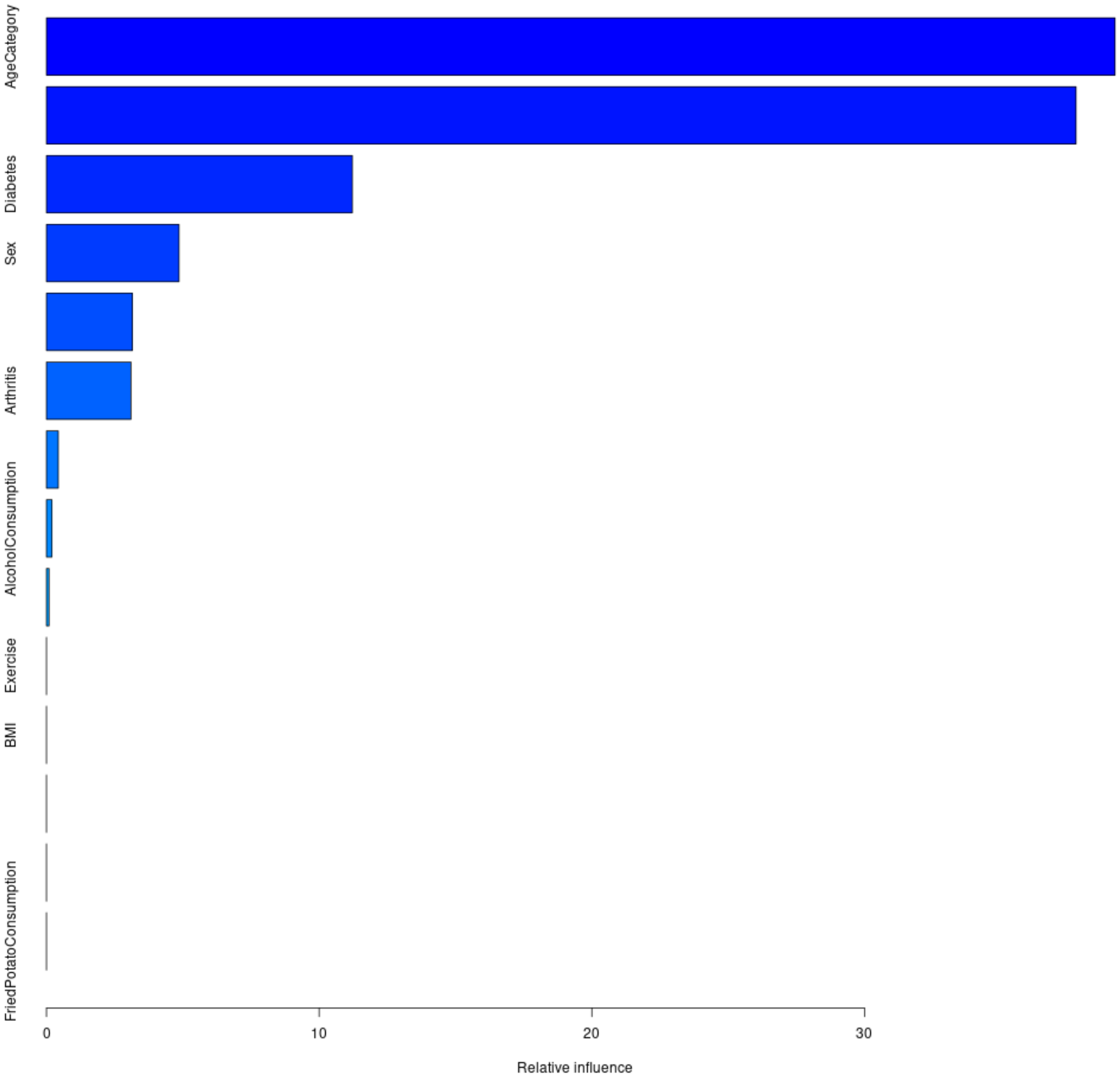

Figure 3: Relative Influence Plot for GBM

Next, we’ll compare the model explanatory capabilities against the remaining 20% of observations:

```
load("~/work/tables/confusion.RData")
conf_matrix
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      1      0
##          1    455    430
##          0  11906 141623
##
##                Accuracy : 0.9201
##                  95% CI : (0.9187, 0.9215)
##     No Information Rate : 0.9199
##     P-Value [Acc > NIR] : 0.4096
##
##                   Kappa : 0.0586
##
##  Mcnemar’s Test P-Value : <2e-16
##
##             Sensitivity : 0.036809
##             Specificity : 0.996973
##          Pos Pred Value : 0.514124
##          Neg Pred Value : 0.922451
##              Prevalence : 0.080051
##          Detection Rate : 0.002947
##    Detection Prevalence : 0.005731
##       Balanced Accuracy : 0.516891
##
##        ’Positive’ Class : 1
##
```

At a cutoff of higher than 50% chance of having heart disease, the specificity is really high but the sensitivity is really low. The degree of agreement (kappa) is also really low. Basically, we can infere that the cut off of 50% is too high to assume certainty about the heart health status for this model. But the model might still be good despite this. Let’s look the ROC curve:

This looks very good! The L bend in the curve shows that the model is very string for prediction. Let’s look at the area under the curve too:

```
load("~/work/tables/model_auc.RData")
model_auc
```

```
## Area under the curve: 0.8315
```

The AUC is really high, we have a good model on our hands despite the lackluster performance at the 50% cutoff.

Conclusion

First, it was surpring to see that diet, exercise and obesity didn’t contribute to predicitng the probability of heart disease as much as we previously thought. One reason might be the cross sectional nature of this data. Future research should look into whether the same results on the variables are replicated. The GBM prediction model that we got from this dataset can also be checked if it also applicable to other datasets and populations; for instance, the BRFSS dataset from 2022.

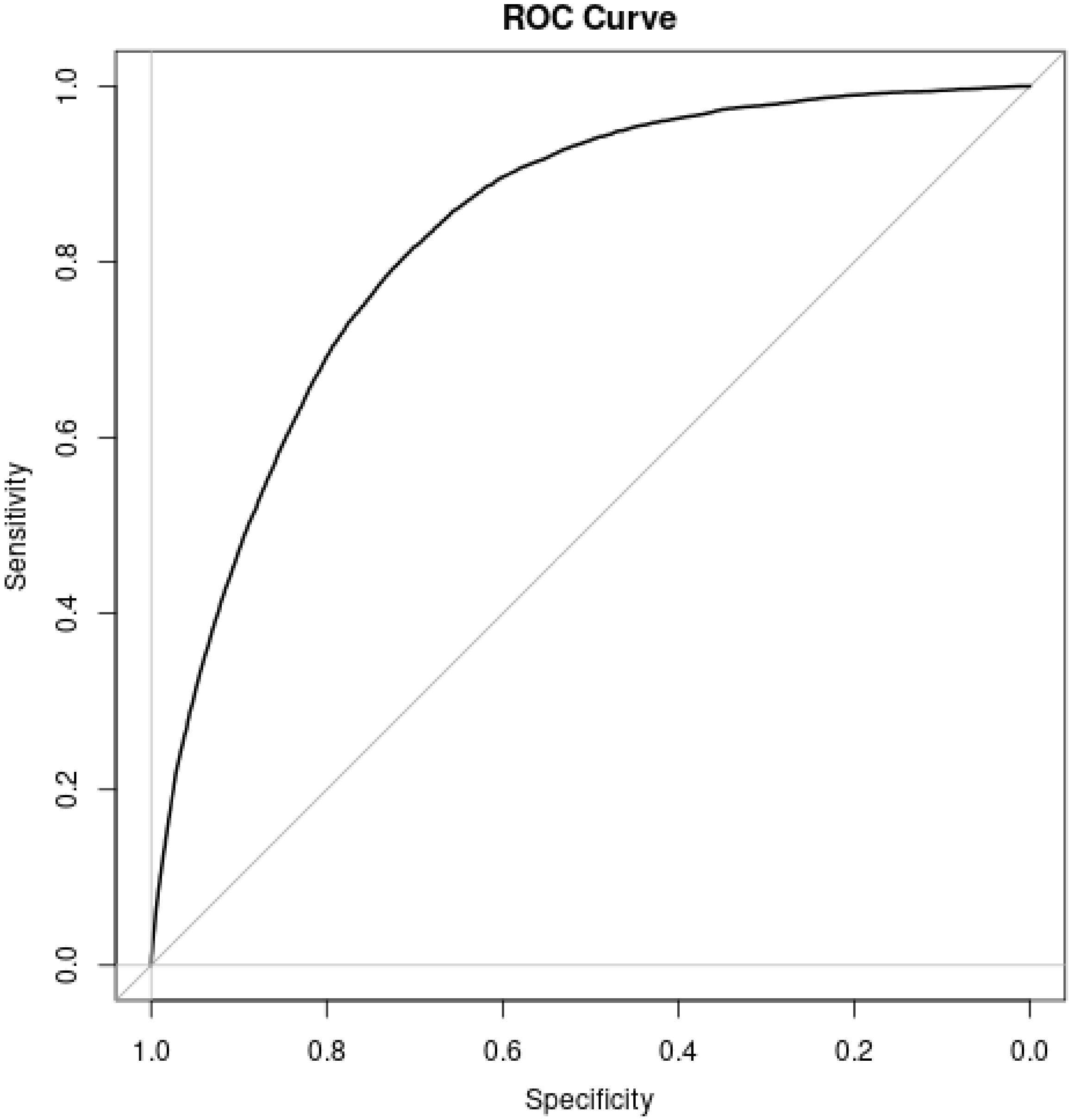

Figure 4: ROC curve for GBM