UK\_Biobank\_PingHsien\_Yang

2023-11-08

# Introduction

## Cholesterol Impact Analysis on Health Outcome Predictive Models

This report presents a comprehensive analysis of a dataset that includes a variety of health-related characteristics of a population, with a particular focus on the impact of cholesterol levels on different health outcomes. The analysis involves statistical summaries, correlations, group comparisons by sex and ethnicity, and advanced modeling using a Gradient Boosting Machine (GBM).

The dataset includes various histograms, bar charts, and scatter plots that outline the distribution of age, sex, ethnicity, education level, Townsend deprivation index, BMI, cholesterol levels, physical activity, smoking and drinking status, and the occurrence of dementia, myocardial infarction (MI), and stroke.

## Objectives

The main objectives of the report are:

To elucidate the distribution and influence of cholesterol levels on the population’s health outcomes. To understand how cholesterol interacts with other variables such as age, sex, BMI, and ethnicity. To evaluate the predictive power of cholesterol levels for outcomes like MI, dementia, and stroke using GBM.

## Variable interpretation

## [1] 10000 13

In this Dataset, we have 10000 Rows and 13 Columns (exclude the ID)

## Statistic summary

## age sex ethnicity\_group   
## Min. :40.0 Length:10000 Min. : 1.000   
## 1st Qu.:50.0 Class :character 1st Qu.: 1.000   
## Median :57.0 Mode :character Median : 1.000   
## Mean :56.3 Mean : 4.667   
## 3rd Qu.:63.0 3rd Qu.: 1.000   
## Max. :70.0 Max. :999.000   
## education\_college\_university\_0 townsend\_deprivation\_index bmi\_0   
## Min. :0.0000 Min. :-6.2583 Min. :12.65   
## 1st Qu.:0.0000 1st Qu.:-3.6912 1st Qu.:24.05   
## Median :0.0000 Median :-2.1967 Median :26.62   
## Mean :0.3388 Mean :-1.4072 Mean :27.29   
## 3rd Qu.:1.0000 3rd Qu.: 0.3075 3rd Qu.:29.62   
## Max. :1.0000 Max. : 9.8924 Max. :68.13   
## cholesterol\_0 MET\_activity smoking\_status\_0 alcohol\_status\_0  
## Min. : 2.074 Min. : 0 Min. :0.0000 Min. :0.000   
## 1st Qu.: 4.917 1st Qu.: 780 1st Qu.:0.0000 1st Qu.:2.000   
## Median : 5.646 Median : 1775 Median :0.0000 Median :2.000   
## Mean : 5.697 Mean : 2641 Mean :0.5816 Mean :1.889   
## 3rd Qu.: 6.441 3rd Qu.: 3546 3rd Qu.:1.0000 3rd Qu.:2.000   
## Max. :10.748 Max. :19278 Max. :9.0000 Max. :9.000   
## dementia\_all\_outcome MI\_all\_outcome stroke\_all\_outcome  
## Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.0144 Mean :0.0531 Mean :0.0323   
## 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000

## Data cleaning - check missing value

## number of missing value : 0

According to the data, there’s no missing values.

## Data cleaning - check error value in categorical variables

## Column: sex   
## Unique Values: M F   
##   
## Column: ethnicity\_group   
## Unique Values: 1 3 4 6 2 5 997 999   
##   
## Column: education\_college\_university\_0   
## Unique Values: 0 1   
##   
## Column: smoking\_status\_0   
## Unique Values: 0 1 2 9   
##   
## Column: alcohol\_status\_0   
## Unique Values: 2 0 1 9   
##   
## Column: dementia\_all\_outcome   
## Unique Values: 0 1   
##   
## Column: MI\_all\_outcome   
## Unique Values: 0 1   
##   
## Column: stroke\_all\_outcome   
## Unique Values: 0 1

According to the data, there’s no error values in categorical variables.

## Data cleaning - check outlinet in Numeric variables

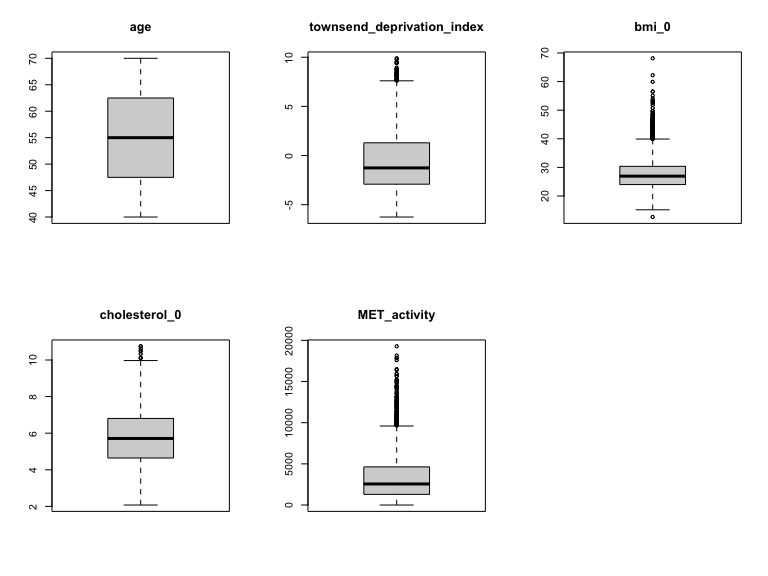


Figure 1 : boxplot for outlier

In the dataset, we detected 3 outliers, which are distributed among “BMI” variables, shows on the chart.

We further investigated the reasons for the existence of outliers, and possible reasons include errors in the data collection process or extreme values.

We Removed outliers to minimize their impact on the results.

## Data cleaning - Remove the Non-Response Data

Ethnicity: remove rows with “Other”, “Prefer not to answer”, and “Do not Know” Smoking: remove rows that is “Prefer not to answer” Alcohol: remove rows that is “Prefer not to answer”

We Removed Non-Response Data to minimize their impact on the results.

## summary statistics after data cleaning

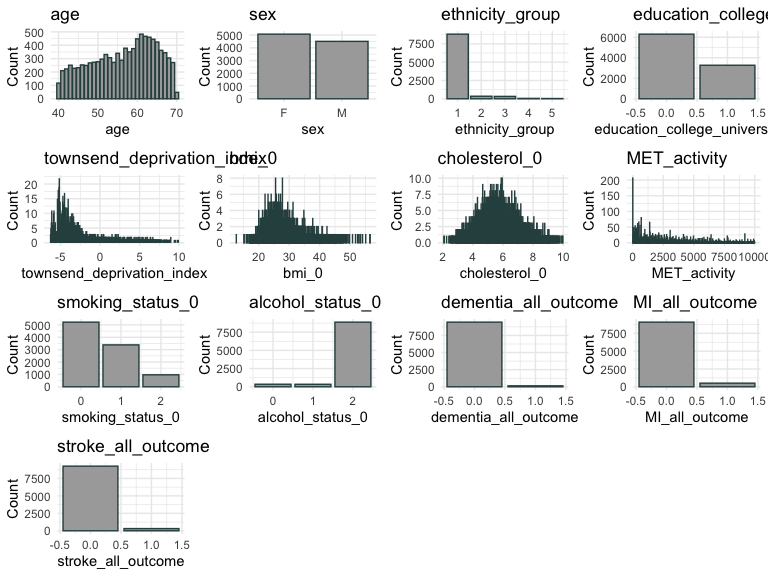


Figure 2 : summary statistic

In this dataset,

**age:** This is a histogram showing the distribution of ages. It looks like a more normal distribution, but may be right-skewed (most individuals are clustered on the lower age side).

**sex:** This is a bar chart showing the distribution of sex in the dataset. There are two categories: female (F) and male (M). As you can see, the number of females in the sample is larger than the number of males.

**ethnicity\_group:** This bar chart shows the distribution of different ethnic groups. Ethnicity 1 (white) is in the majority.

**education\_college\_university:** This histogram appears to show the distribution of college-educated people with a larger number of non-college professors.

**townsend\_deprivation\_index:** This is a histogram showing the distribution of the Townsend Deprivation Index. This index may be a measure of the socio-economic status of the area. It is left-skewed (most individuals are concentrated on the lower side of the index).

**bmi\_0:** This histogram shows the distribution of body mass index (BMI). The data appears to be more normally distributed and concentrated within a certain interval.

**cholesterol\_0:** This is a plot of the distribution of blood cholesterol levels. The data appears to be close to normally distributed, perhaps slightly skewed to the right.

**MET\_activity:** This histogram represents a measure of exercise. The distribution shows a lot of small peaks, which may represent a breakdown of different activity levels.

**smoking\_status\_0:** This bar graph shows the distribution of smoking status, coded with numbers to indicate different smoking statuses.

**alcohol\_status\_0:** This is another bar chart showing the distribution of drinking status, again coded with numbers.

**dementia\_all\_outcome:** This bar chart shows the distribution of dementia outcomes, with 0 and 1 indicating the absence and presence of dementia.

**MI\_all\_outcome:** This bar graph shows the distribution of all myocardial infarction outcomes, with 0 and 1 indicating the absence and presence of myocardial infarction.

**stroke\_all\_outcome:** This bar chart shows the distribution of all stroke outcomes.

### Marginal Correlation and Correlation martix

### Correlation Heatmap

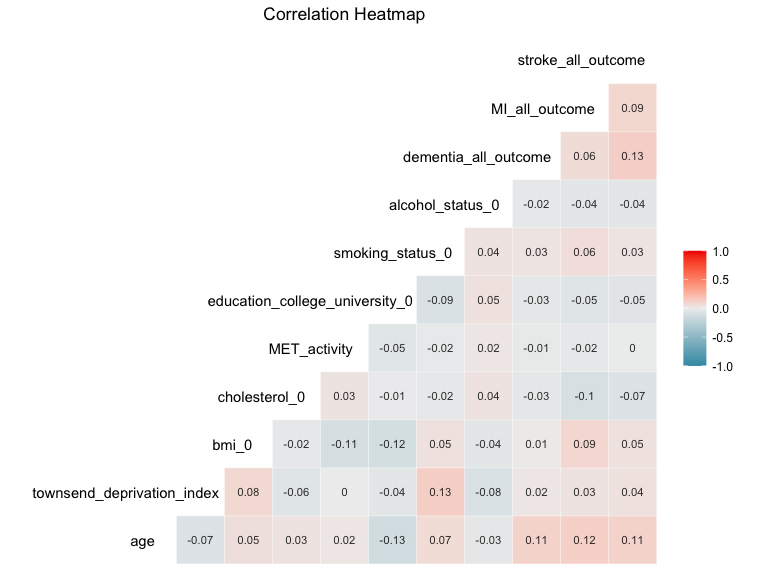


Figure 3 : Correlation Heatmap

The heatmap shows that most variables have a very low correlation with each other since the majority of the values are close to 0. However, there are some notable correlations, such as:

1.A moderate **positive** correlation (0.13) between **dementia\_all\_outcome** and **stroke\_all\_outcome**, suggesting that individuals who have had a stroke might be more likely to be diagnosed with dementia.

2.A small **positive** correlation (0.11, 0.12) between **age**, **dementia\_all\_outcome**, **MI\_all\_outcome** and **stroke\_all\_outcome**, implying that the likelihood of having a heart attack or stroke may increase with age.

3.A small **negative** correlation (-0.1) between **cholesterol** and **MI\_all\_outcome**,This means that with a slight decrease in cholesterol levels, there was a slight increase in the incidence of all outcomes of heart attacks.

4.A small **positive** correlation (0.09) between **BMI** and **MI\_all\_outcome**, implying that the likelihood of having a heart attack with higher BMI.

5.A small **positive** correlation (0.08) between **townsend\_deprivation\_index** and both **BMI** and **Smoke\_status**, suggesting that as deprivation increases, there tends to be a slight increase in BMI and a higher likelihood of smoking. But a small**negative** correlation (-0.08) with **alcohol\_status**, where higher deprivation scores are associated with a slight decrease in alcohol\_status.

However, those associations are very weak, suggesting that there is no strong direct relationship between each variables. Those slight correlation between each are not sufficient to indicate a clear trend or to inform clinical decisions. More comprehensive studies are needed to further investigate the relationship between these variables.

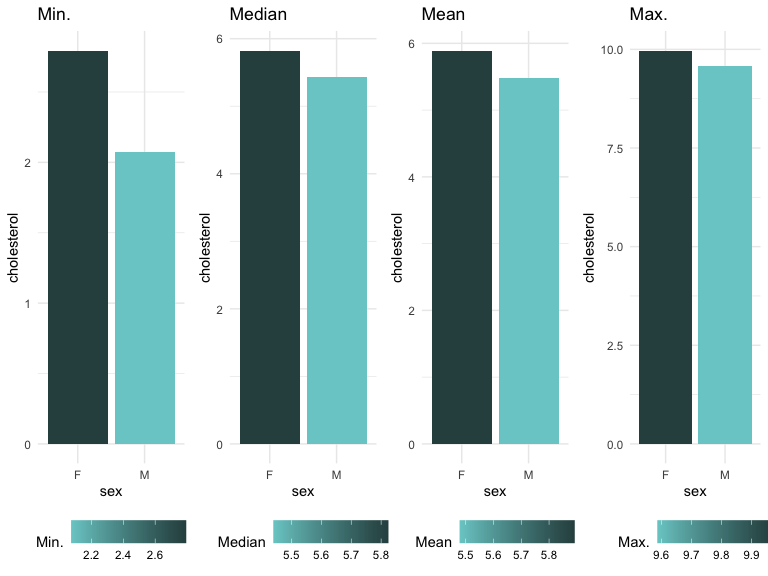


Figure 4 : gendar choleserol comparison

This image presents a set of four bar charts, each one representing a different statistical measure (Minimum, Median, Mean, and Maximum) of cholesterol levels across Male and Female. The cholesterol levels are on the vertical axes, and the Sex are on the horizontal axes.

As can be seen from the bar graphs, females have slightly higher cholesterol than males. However, Males and Females in this particular dataset are very close to each other in terms of minimum, median, mean, and maximum cholesterol values. This may mean that gender has little effect on cholesterol levels, but more analysis is needed to support this conclusion, considering standard deviation and other distributional characteristics.

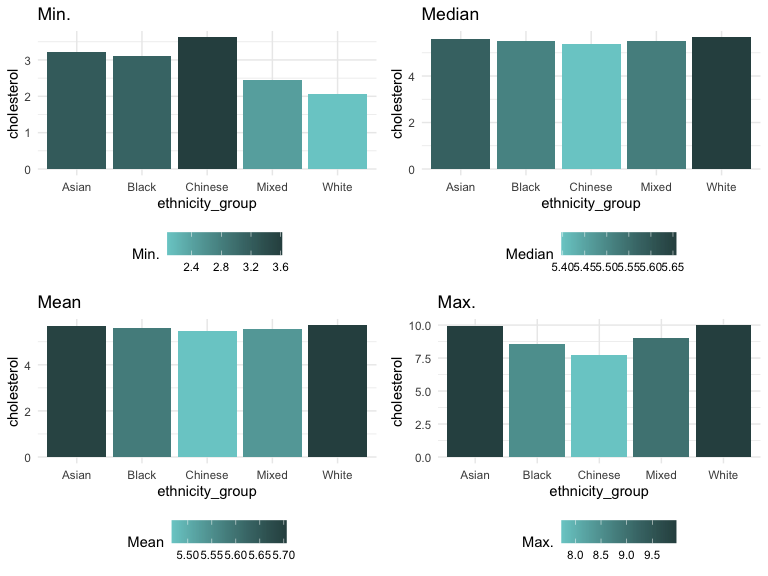


Figure 5 : ethnicity cholesterol comparison

This image presents a set of four bar charts, each one representing a different statistical measure (Minimum, Median, Mean, and Maximum) of cholesterol levels across five different ethnicity groups (Asian, Black, Chinese, Mixed, and White). The cholesterol levels are on the vertical axes, and the ethnicity groups are on the horizontal axes.

As can be seen from the bar graphs,The lowest minimum value of cholesterol is seen in the Mixed ethnicity group. The highest maximum values of cholesterol are seen in the Asian and White ethnicity groups, both just below 9.5, while the lowest maximum is in the Chinese ethnicity group, around 8.0.

The uniformity of median and mean levels suggests that there’s not a significant difference in central cholesterol levels among these ethnic groups, but there may be differences in the range (minimum and maximum), which could be important for medical or health-related studies.

### Checking whether cholesterol\_0 is normal distributed ot not

##   
## Asymptotic one-sample Kolmogorov-Smirnov test  
##   
## data: BM\_ml$cholesterol\_0  
## D = 0.018306, p-value = 0.003266  
## alternative hypothesis: two-sided

The output provided is from an Asymptotic one-sample Kolmogorov-Smirnov (K-S) test. This test is a nonparametric test used to determine whether a sample comes from a specified distribution.

p-value = 0.003266: The p-value indicates the probability of observing a test statistic as extreme as, or more extreme than, the one observed if the null hypothesis is true. In this case, the null hypothesis is that the sample distribution is the same as the reference distribution. The p-value is 0.003266, which is less than 0.05, suggesting that there is a statistically significant difference between the distribution of cholesterol\_0 and the reference distribution, which means the *cholesterol\_0 variable is not normal distributed*.

### split dataset

### using Box–Cox transformation

##   
## Asymptotic one-sample Kolmogorov-Smirnov test  
##   
## data: BM\_training$cholesterol\_0  
## D = 0.011162, p-value = 0.3737  
## alternative hypothesis: two-sided

Use Box-Cox transformations to improve the normality of data.

Find the value of λ that corresponds to the largest y-value (likelihood value) in the likelihood-ratio test, which is the value of λ that we believe best transforms the data into a normal distribution. If the best λ-value is 0, the natural logarithmic transformation is used; if it is not 0, the transformation is performed using a more complicated formula.

Then use statistical tests after the transformation to test the effect of the improvement. This will improve the accuracy and predictive power of the model.

### Compare different model

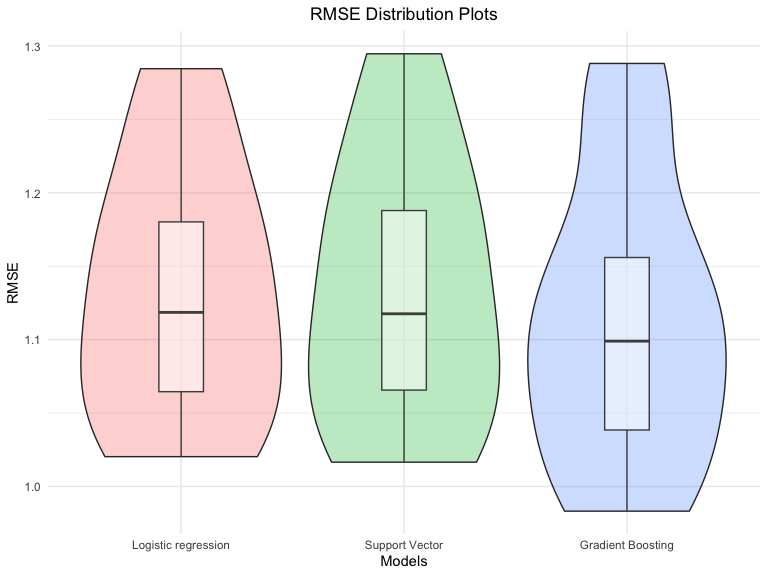


Figure 6 : Model RMSE comparison

This graph shows the distribution of RMSE (Root Mean Square Error) for three different statistical models. In forecasting models, RMSE is a common measure of the difference between the model’s predicted value and the actual observed value. Generally speaking, the lower the RMSE, the higher the accuracy of the model.

The three models are:

1.**Logistic regression** 2.**Support Vector Models** 3.**Gradient Boosting**

As can be seen from the violin plot, **Gradient Boosting** has lowest RMSE, which means it has better predictive accuracy. Therefore, for further analysis and model interpretation, I will choose the *Gradient Boosting model*, taking into account the stability and high accuracy of its prediction results.

### cholesterol\_0 with tuning

## Mean Absolute Percentage Error : 16.22%

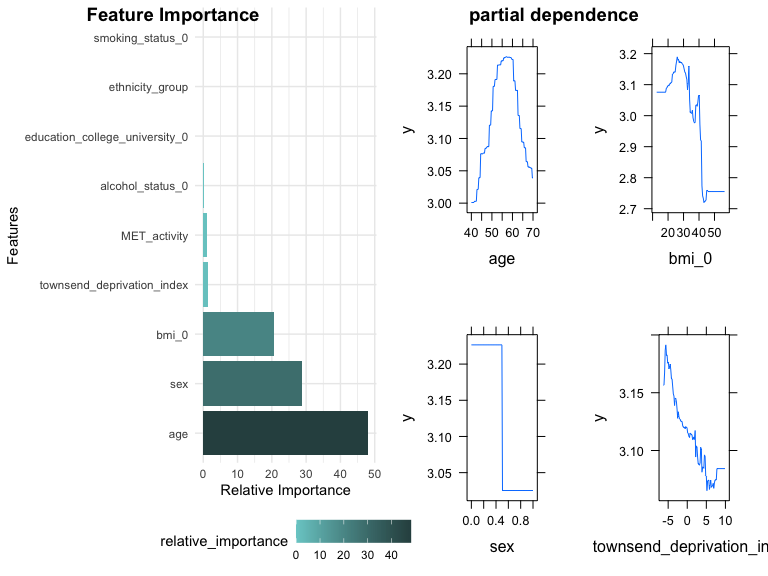


Figure 7 : cholesterol model interpretatIon

Feature Importance Bar Chart :

It shows the relative importance of individual features to model predictions. Feature importance is a way of measuring the magnitude of the role that features play in the model’s decision-making process. In this graph, ‘**age**’ (age) is the most important feature, followed by ‘**bmi\_0**’ (body mass index) , ‘**sex**’ (gender) and ‘**townsend\_deprivation\_index.**(deprivation index)’ . This means that these variables have the most influence in the model predictions.

partial dependence plots :

It shows the relationship between the six most important characteristics and the target variable Y. The graph shows the relationship between the six most important characteristics and the target variable Y. In those chart, ‘**age**’: we can see that the mean of Y starts to decrease after peaking at a certain age point,This suggests that the relationship between age and Y is important in that particular age range, which is at **55** to **60**.

‘**bmi\_0**’: This may point to a correlation between high BMI and an increase in Y(25 to 30), but after a certain point, the relationship becomes weaker.

‘**gender**’: Since gender is a categorical variable, what we see is that different genders may have different average effects on the Y-value . we can see that female(0) has more impact than male(1) in cholesterol level.

‘**Townsend\_deprivation\_index**’: negative correlate, wiht higher Townsend\_deprivation\_index, we can see that cholesterol correlation became weaker.

Finally, A MAPE of 15.58% means that the model’s predicted values will be in error by 15.58% on average.

### predict the disease by gbm model

### MI

## Using 300 trees...

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Threshold corresponding to the largest Youden's index : 0.0481002

##   
## Accuracy: 68.78 %

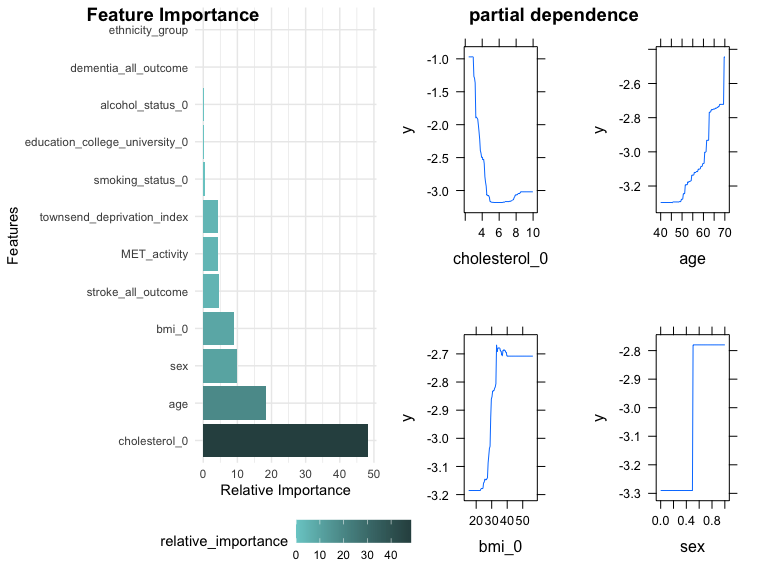


Figure 8 : MI model interpretatIon

Because MI\_all\_outcome is a categorical variable, i.e., its values are only 0 and 1. Then when modeling with GBM (Gradient Boosting Machine), we should use the binomial distribution (“bernoulli”) method.

We used **Youden’s J statistic** to choose threshold for prediction, This is a statistical method used to select a threshold that maximizes the difference between the true and false positive rates to find a good balance.

Feature Importance Bar Chart :

In this graph, ‘**cholesterol\_0**’ (cholesterol) is the most important feature, followed by ‘**age**’ (age) , ‘**bmi\_0**’ (BMI) and ‘**sex**’(gender). This means that these variables have the most influence in the model predictions, especially **cholesterol** that occupy 50%.

partial dependence plots :

**cholesterol** : The graph shows a rapid decline in the predicted value of Y as the cholesterol level increases from the lowest value to about 4 units, and then a stabilization of the predicted value of Y as the cholesterol level increases further. This may mean that cholesterol levels have a strong negative effect on Y within a certain range, but once this range is exceeded, the effect becomes less pronounced.

**age** : As age increases, the predicted value of Y shows a gradual increase, especially at around 60 years of age, where the predicted value increases more significantly. This suggests that age may be a factor in the increase in Y, especially at older ages.

**bmi\_0** : For BMI, the predicted value of Y has a spike at a BMI of about 30 units and then stabilizes. This may indicate that there is a significant association between BMI and Y in the middle range of the model, but that this association no longer strengthens once BMI reaches a certain point.

**sex** : For gender, we see two different levels, indicating that males and females have significantly different predicted impacts on Y in the model. We can say that male(1) has more chance to get myocardial infarction.

Finally, this model has approximate 70 % Accuracy.

### Dementia

## Using 222 trees...

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Threshold corresponding to the largest Youden's index : 0.01886553

##   
## Accuracy: 77.25 %

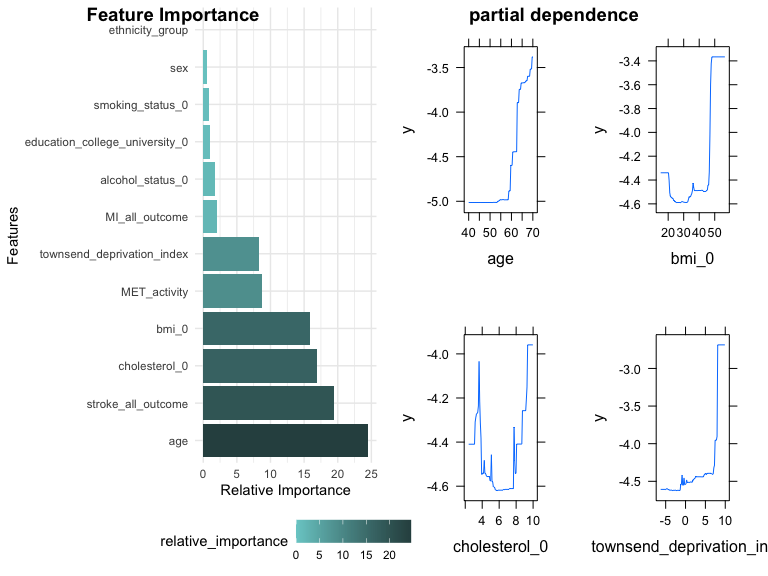


Figure 9 : Dementia model interpretatIon

Same as MI\_all\_outcome, dementia\_all\_outcome is a categorical variable. we use the binomial distribution (“bernoulli”) method for modeling.

We used **Youden’s J statistic** to choose threshold for prediction, This is a statistical method used to select a threshold that maximizes the difference between the true and false positive rates to find a good balance.

Feature Importance Bar Chart :

In this graph, ‘**age**’ (age) is the most important feature, followed by , ‘**bmi\_0**’ (BMI), ‘**cholesterol\_0**’ (cholesterol) and ‘**townsend\_deprivation\_index**’(deprivation index). This means that these variables have the most influence in the model predictions, especially **age**.

partial dependence plots :

**age** : As age increases, the predicted value of Y shows a gradual increase, especially at around 60 years of age, where the predicted value increases more significantly. This suggests that age may be a factor in the increase in Y, especially at older ages.

**bmi\_0** : It shows an interesting non-linear relationship where the predicted value of Y decreases in certain BMI ranges, especially in the range of about 30. This means that in this range, the effect of increasing BMI on Y is negative. In other ranges, however, the relationship appears to become positive or insignificant.

**cholesterol\_0** : In the graph of cholesterol levels, we see several sharp peaks(3.5 , 8) and valleys(6 to 8) that show significant changes in the predicted value of Y at specific cholesterol levels. This indicates that the model is very sensitive to changes in cholesterol levels, especially at certain specific points, which may mean that there is a specific effect of cholesterol levels on the predicted variables around these points

**townsend\_deprivation\_index** : For the Townsend Deprivation Index, the graph shows a complex nonlinear relationship. The predicted value of Y appears to be lowest when the index is near 0, while the predicted value increases at higher or lower values of the deprivation index. This may indicate that deprivation has a smaller effect on Y in the medium range and a larger effect at higher levels of deprivation.

Finally, this model has approximate 70 % Accuracy.

### Stroke

## Using 277 trees...

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Threshold corresponding to the largest Youden's index : 0.02614842

##   
## Accuracy: 66.97 %

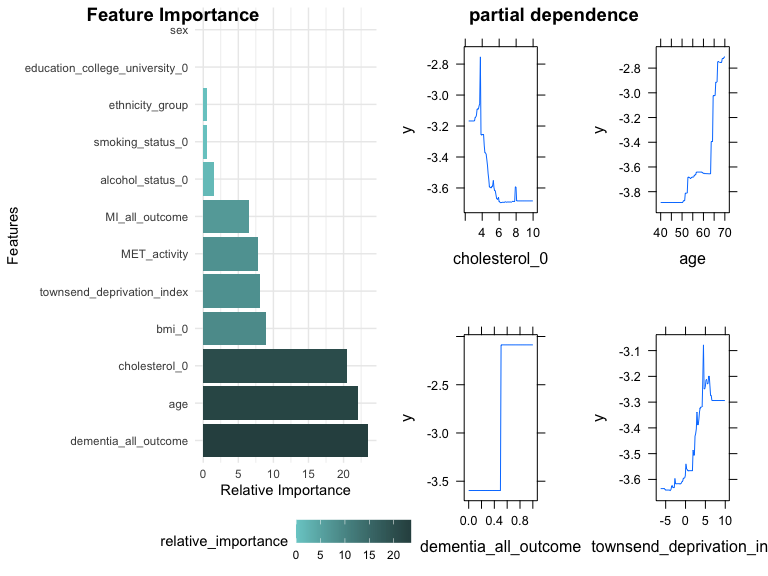


Figure 10 : Stroke model interpretatIon

Same as MI\_all\_outcome, stroke\_all\_outcome is a categorical variable. we use the binomial distribution (“bernoulli”) method for modeling.

We used **Youden’s J statistic** to choose threshold for prediction, This is a statistical method used to select a threshold that maximizes the difference between the true and false positive rates to find a good balance.

Feature Importance Bar Chart :

In this graph, ‘**age**’ (age) is the most important feature, followed by , ‘**cholesterol\_0**’ (cholesterol), ‘**dementia\_all\_outcomeand**’(dementia) and ‘\*MET”\_acticity\*’(measure for exercise). This means that these variables have the most influence in the model predictions.

partial dependence plots :

**age** : This graph shows that as age increases, the predicted value of Y shows an increasing trend, especially in the 60-70 age range, where this trend is very pronounced. This may mean that the effect of age on the predicted outcome increases at that age, which may indicate a positive correlation between age and the predicted target.

**Cholesterol\_0** :In the dependency plot for cholesterol levels, it can be observed that the predicted value of Y decreases significantly when cholesterol levels are in the range of specific low values. However, for most of the range of cholesterol levels, the value of Y is relatively stable, suggesting that it only has a significant impact on the predicted target at specific low cholesterol levels.

**Townsend\_deprivation\_index** : This graph shows a significant peak in the effect of the Thomson Deprivation Index on the predicted value of Y, which occurs when the index is close to 0. This may indicate that the predicted value of Y is lowest at moderate levels of deprivation and increases at higher or lower levels of deprivation. This may indicate that the predicted value of Y is lowest at moderate levels of deprivation and increases at higher or lower levels of deprivation.

**MET\_activity** : The graph of MET activity levels shows that the predicted value of Y is relatively stable at low activity levels, with a slight increase in the range from 2000 to about 6000. However, at very low activity levels or close to 10,000, the predicted value of Y increase significantly. This may indicate that only at very high or very low activity levels does the MET activity level have a significant effect on the predicted results.

Finally, this model has approximate 70 % Accuracy.

# Conclusion

Based on the provided data and model results, we can summarize the effect of cholesterol (Cholesterol) in different disease prediction models as follows:

**Cholesterol Distribution:**

Appears roughly normal with slight right-skewness. No substantial difference in cholesterol levels when comparing gender. Variance is seen across ethnic groups, with Asian and White ethnicities displaying the highest maximum cholesterol levels.

### Effect of Cholesterol on Myocardial Infarction (MI) Modeling:

Cholesterol is one of the most important features in the Myocardial Infarction (MI) prediction model, especially since it accounts for 50% of the feature importance. From the Partial Dependence Plots, it can be observed that the predictive value Y decreases rapidly as the cholesterol level increases from the lowest value to approximately 4 units, and then stabilizes as the cholesterol level increases further. This may indicate that cholesterol levels have a strong negative effect on the risk of myocardial infarction within a certain range, but that the effect diminishes beyond this range.

### Effect of cholesterol on the cognitive disorder (Dementia) model:

In models of cognitive disorders, cholesterol is less important than age and BMI, but still has a place among the important characteristics. The dependency plots show spikes and troughs at specific cholesterol level points (3.5 and 8 units), suggesting that the model is very sensitive to specific points in cholesterol levels, and may imply that cholesterol levels around these points have a specific effect on the predicted variables.

### Effect of cholesterol on stroke (Stroke) modeling:

**Cholesterol** is also an important feature in stroke prediction models. The dependency plot shows that when cholesterol levels are within a specific range of low values, the predicted value of Y decreases significantly. However, Y was relatively stable over most of the range of cholesterol levels, suggesting that cholesterol levels have a significant effect on the predictive target only within a specific low range of values(2 to 4).

In summary, **cholesterol** significantly affected predictive outcomes in the cardiovascular disease (e.g., myocardial infarction and stroke) prediction models, particularly in the low to moderate range. However, in models of cognitive disorders, cholesterol appears to have a smaller effect than other characteristics such as age and socioeconomic status. Despite the differences in the predictive contribution of cholesterol to these diseases, maintaining healthy cholesterol levels remains important in the prevention and management of these diseases.

To sum up, cholesterol levels are notably influential in the prognostic models for cardiovascular conditions such as myocardial infarction and stroke, particularly within the low to moderate spectrum (**2 to 4 units**). It is imperative that clinical attention is heightened when patient cholesterol levels fall within this critical range, as it may significantly alter the predictive outcomes and thereby inform more targeted intervention strategies.

However, in models of brain dysfunction, the effect of **cholesterol** appears to be smaller than that of other characteristics such as **age** and **socioeconomic** status. Despite the differences in the predictive contribution of cholesterol to these diseases, maintaining healthy cholesterol levels remains important in the prevention and management of these diseases.

## Recommendations for Further Analysis:

**Deeper Statistical Analysis:** To clarify the negative correlation between cholesterol and MI outcomes.

**Expanded Data Collection:** To better understand cholesterol’s role across different demographics and health profiles.