Heart Problem: Xtreme Gradient Boosting

STATISTCAL CONSUTLING REPORT

Venkatesh Manikantan
KU LEUVEN |

Data Exploration:

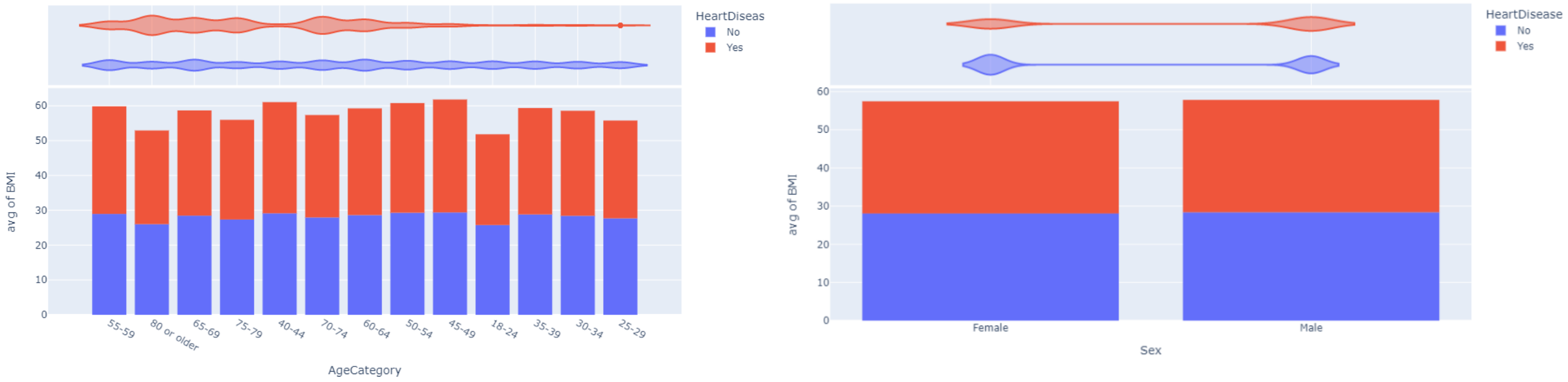

Fig 1: Age Category Bin with Avg BMI (Right), Gender with Avg BMI (Left)

The dataset consists of 17 feature variable and one target variable i.e., an indicator variable which showcases presence of heart disease. Out of the 17 variables, 4 of the variables are numerical in nature and rest of the variables are categorical.

Moreover, to be able to use categorical data in our boosted tree model, a few encoding methods were employed such as dummy, ordinal and weight of evidence encoding. Dummy encoding was used on variables which were binary such as sex, ordinal encoding was used in case where the categorical data showcased a nature of magnitude, such as general health, which has a range from poor to very good status.

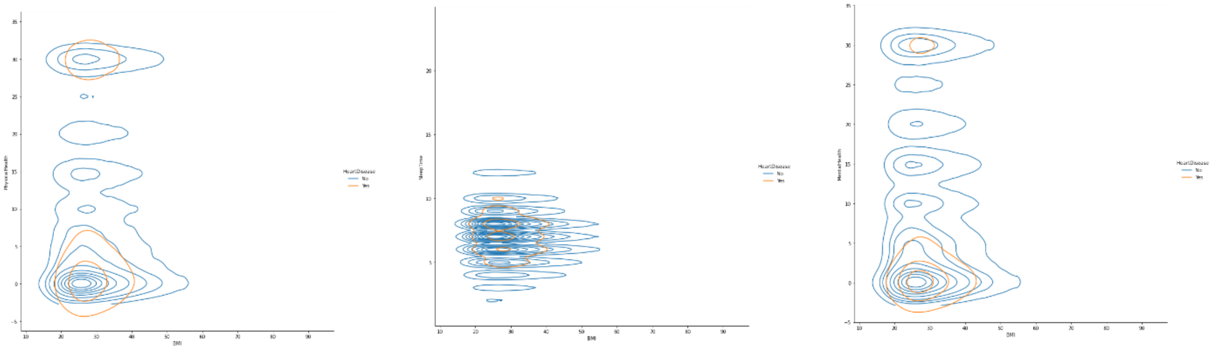

Fig 2: Distribution Plot BMI vs Physical Health, BMI Vs Sleep and BMI Vs Mental Health (In order)

From the above distribution plots, it can be observed that most of the observations consist of heart disease have lower physical and mental health scores. On the other hand, the observations tend to be more evenly spread across both categories when it come to BMI and sleep time in comparison to the other two variables.

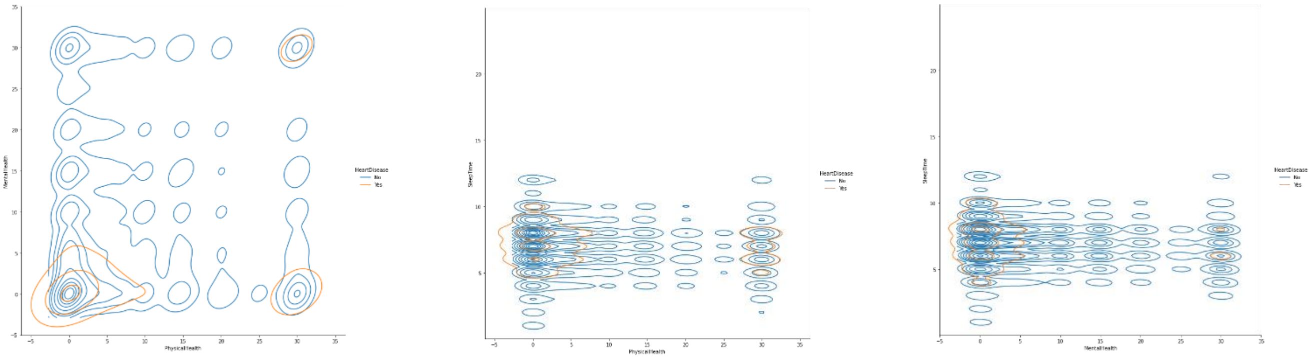

Fig 3: Distribution Plot Physical Vs Mental, Physical Vs Sleep Time and Mental Vs Sleep time (In Order)

Figure 3, just gives another perspective, where we can observe that most of the observations who have heart disease come right in the range of to 5 – 10 in sleep time, with respect to the other two variables where most of the observations who have heart disease tend to have a lower score in both mental and physical health. Interesting, point to note is that there are a few observations where heart disease is present where they have high physical score but a low mental score and even fewer observations where the person tend to have high score in both physical and mental health.

PCA:

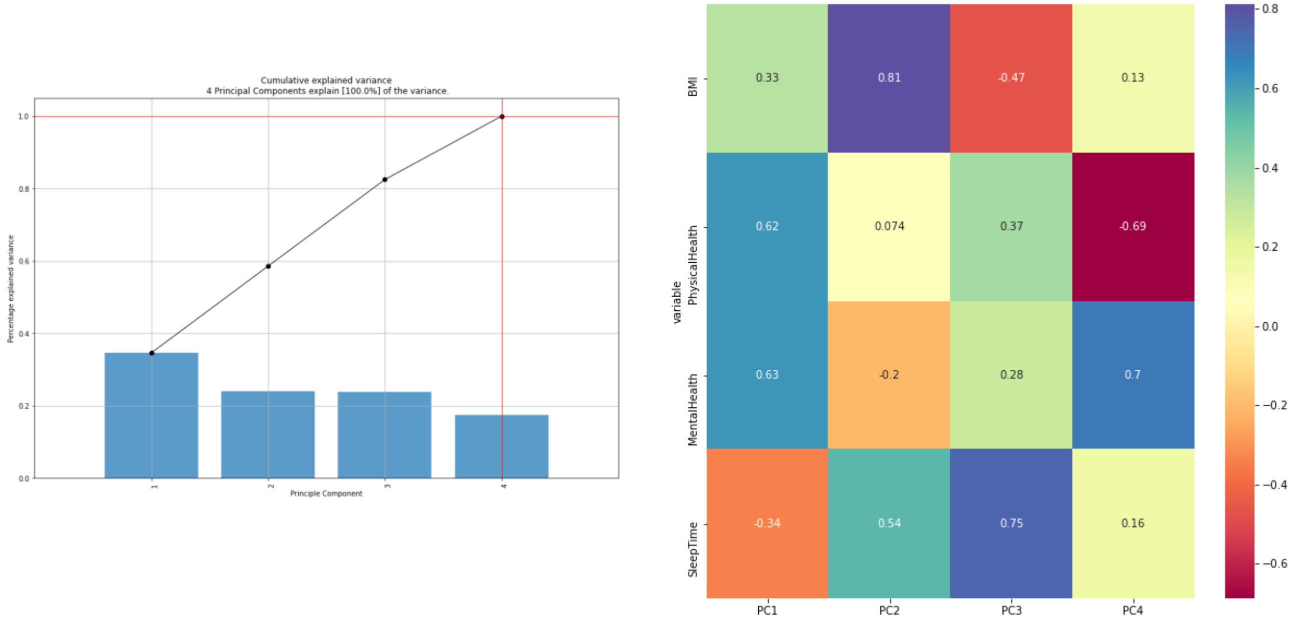

Fig 3: PCA scree plot (Right) , PCA Loading Plot (Left)

From the PCA scree plot, it can be seen that 100% of the variance is explained by just 4 principal components and therefore we can look into the first 3 principal components and understand how the numerical variables relate to one another.

From the first PC, it can be observed that both physical health and mental health tend to slightly group up together with loading value of 0.62 and 0.63 respectively. On the other hand, both BMI and sleep time are significant loading of greater than 0.75 on the 2nd and 3rd principal component respectively.

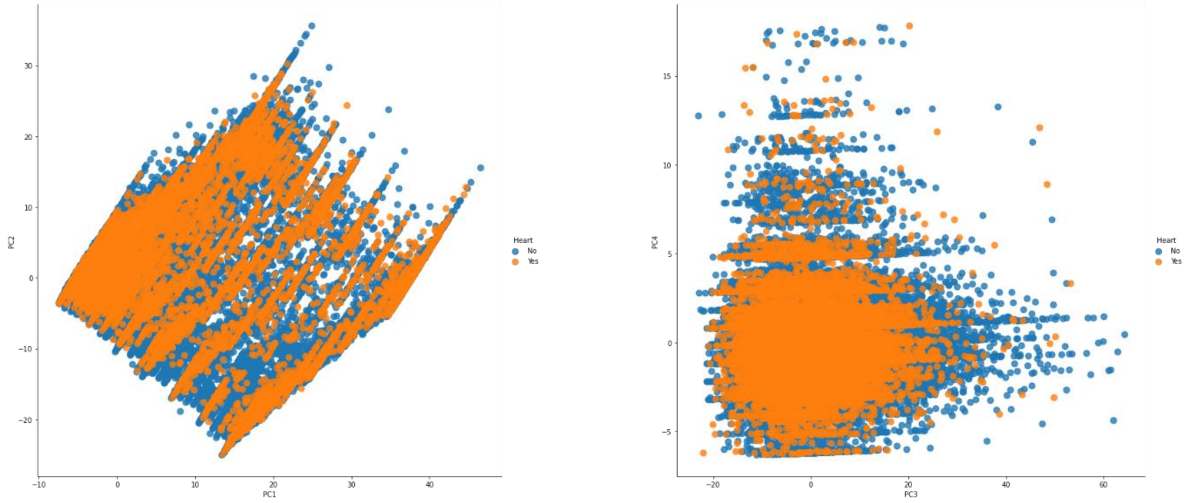

Fig 4: PC1 vs PC2 (Hue = “heart disease”), PC3 vs PC4 (Hue = “heart disease”)

The idea behind plotting, these figures is to observe if there is a decision boundary separating the binary variable, heart disease. In this case it seems like, there is not a clear boundary when we use a linear compression technique such as PCA. A kernel PCA can also be observed before jumping on to building a boosted tree model to further analyze if there is a nonlinear decision boundary between the two classes

Over Sampling:

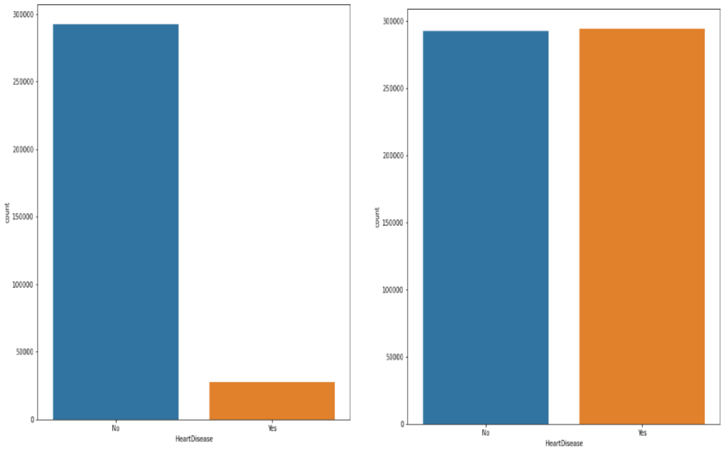

Fig 5: Count of Heart Disease Presence before and after Oversampling

Techniques such as oversampling are mandatory, when there is such a heavy class imbalance. Building a decision tree model without correcting for the class imbalance would lead to very poor performing model in detecting heart disease as the model in many cases, would not have the observations where heart disease is present. And will also falsely lead to a higher accuracy score as most the observation would be classed as the one belonging to the majority class which is in this context lead to no presence of heart disease.

XGBOOST MODEL VALIDATION:

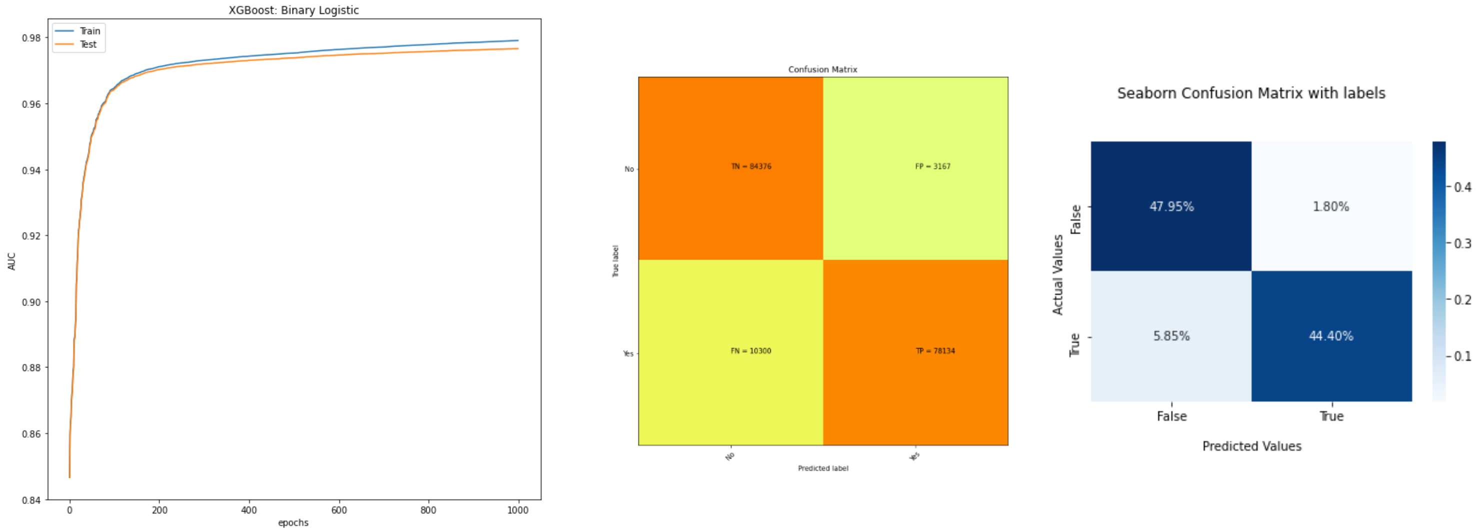

Fig 6: AUC plot with respect to number of iterated boosted trees, Confusion Matrix with Magnitude and Confusion Matrix with Percentage

AUC plots tend to take into account of both true negatives and false negatives and the training set AUC is reaching about 0.98 and the testing set is just a bit lower than the training set. Moreover, from the confusion matrix we can observe the misclassification error in total is only about 7.65 %, keeping in mind that we have corrected for imbalance using over sampling, this model can be relied upon when it is classifying between both the classes and the features that tend to be important for a model with 92.35% accuracy is worth looking into.

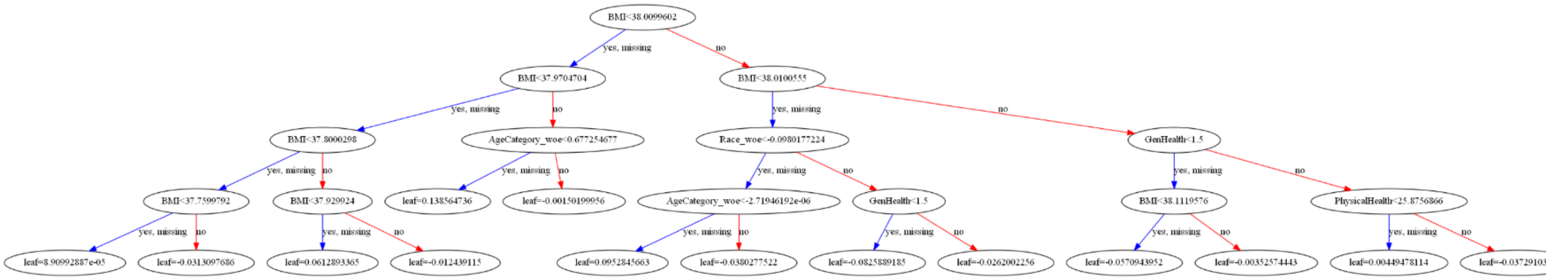

Fig 7: The Final Decision Tree with good performance by using early stopping to avoid overfitting

Plotting one of the good performing decision tree from the boosted tree iterations is a good initial sign of looking into the variables which are selected as the root by the decision tree. From observing the figure, it can be seen that the model favors BMI to classify between the two classes more than other variables, as it can be seen to showcase repeatedly even in a such a low depth tree of only four. Moreover, categories such as Race, Age Category, General Health and Physical Health also show up in this decision tree. Therefore, from our primary analysis we can see that these 5 variables tend to be more important while constructing the decision tree to classify between the presence and absence of heart disease.

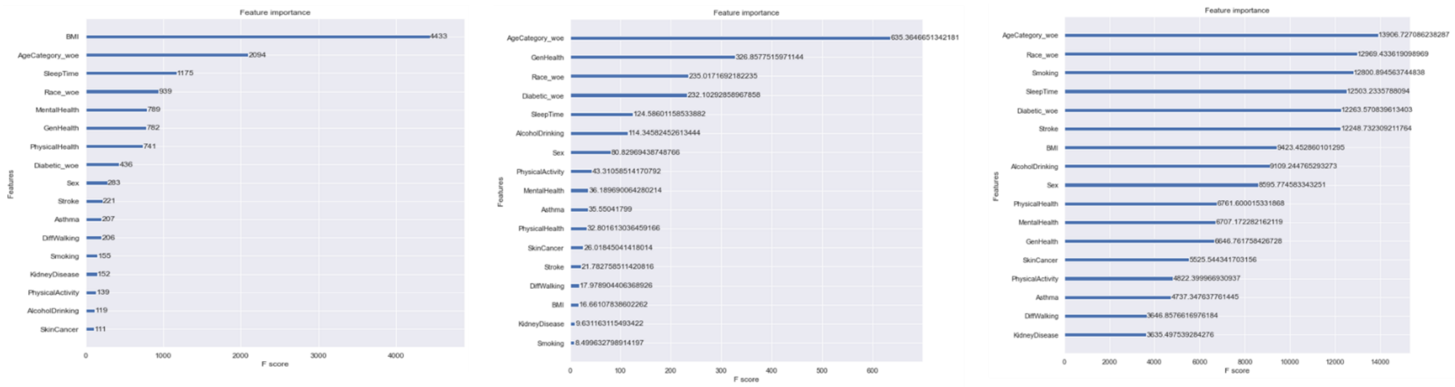

Fig 8: Feature Importance with respect to Weight, Gain and Cover in that order respectively

From the three plots we can gain further insights on the importance of the variables, from the weight plot, which represents how many times a variable has been used by the model shows that BMI , Age , Sleep time, Race ,Mental Health and General health are the top 6 variables which are used repeatedly by the model.

On the other hand, form the gain plot which showcase variable importance with respect to how they help the model reduce error across various iteration shows a different set of variables, which are Age, General health, Race, Diabatic and Sleep time as top 5 variables which helps the model learn and reduce the errors substantially. Interesting to note that.

Finally, from the cover plot which is the weighted representation of the variables used by the model with respect to how many data points pass through those variables. It can be once again observed that Age, Race, Smoking, Sleep Time and Diabetic are the top 5 important variables.

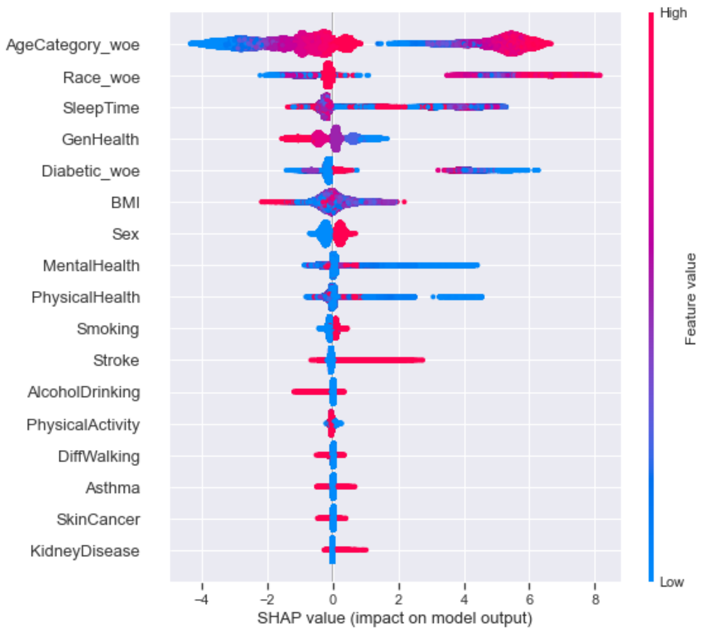

Fig 9: SHAP Feature Importance Plot

From figure 9, it can be noticed that the overall feature importance ranking, is as follows:

- Age
- Race
- Sleep Time
- General Health
- Diabetic
- BMI

Other interesting things of note is that the model is showcasing that there is an even split between genders it seems the model seems to favor men to be more prone to heart disease. And same can also can be seen for smoking as when the dummy variable shows, the model favors it be a problem with heart disease. Furthermore, Age and Race are divided into two blocks, which can only be analyzed by decoding the encoding performed on them.

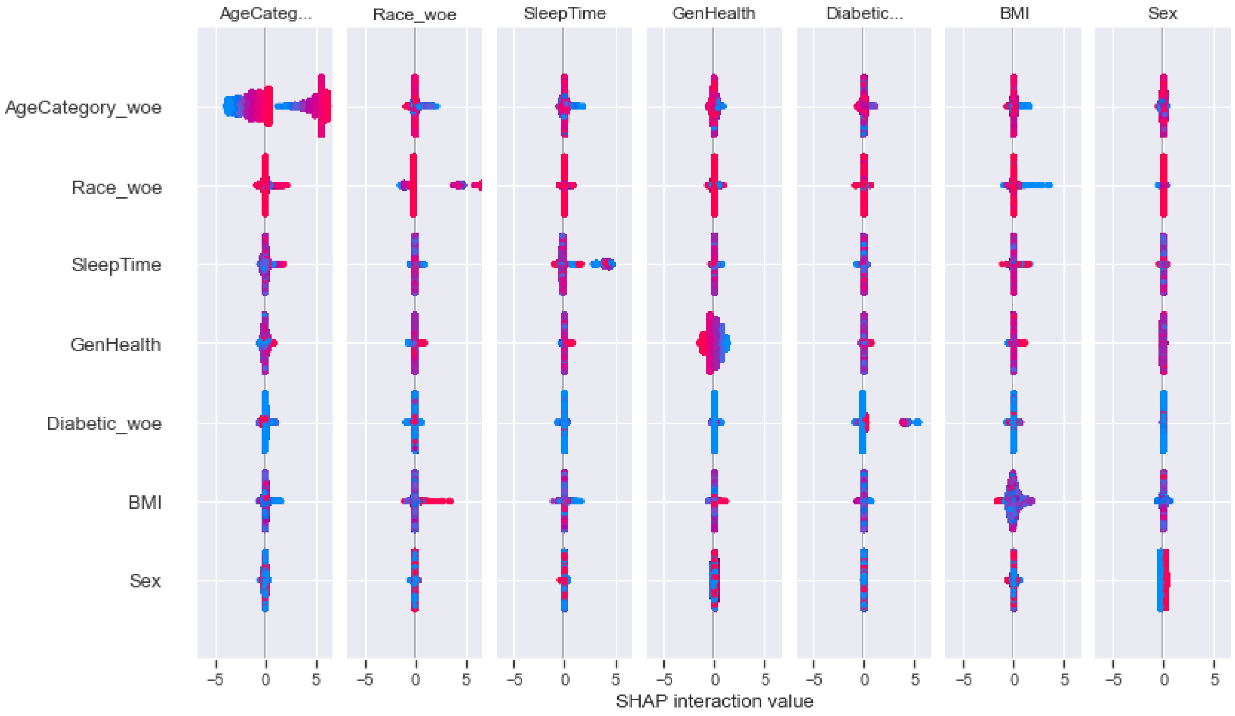

Fig 10: SHAP Interaction Plot

From the interaction plot it can be observed that just one variable has slight interactions with other variable and this variable is BMI, which has the most interaction with Race, followed by sleep time and general health.

And even though BMI doesn’t show up in the final top 5 variable, it is a variable which sure can’t be ignored

Conclusion:

Given that the model has a high performance of AUC greater than 0.97 it is safe to gather inference from the feature importance measure. And it can be concluded from all the findings that Age_Category variable is one the most deciding factors in classifying between the two classes, closely followed by Race, Sleep Time, these three variables seem to be the most important to determine if a person would have a heart disease