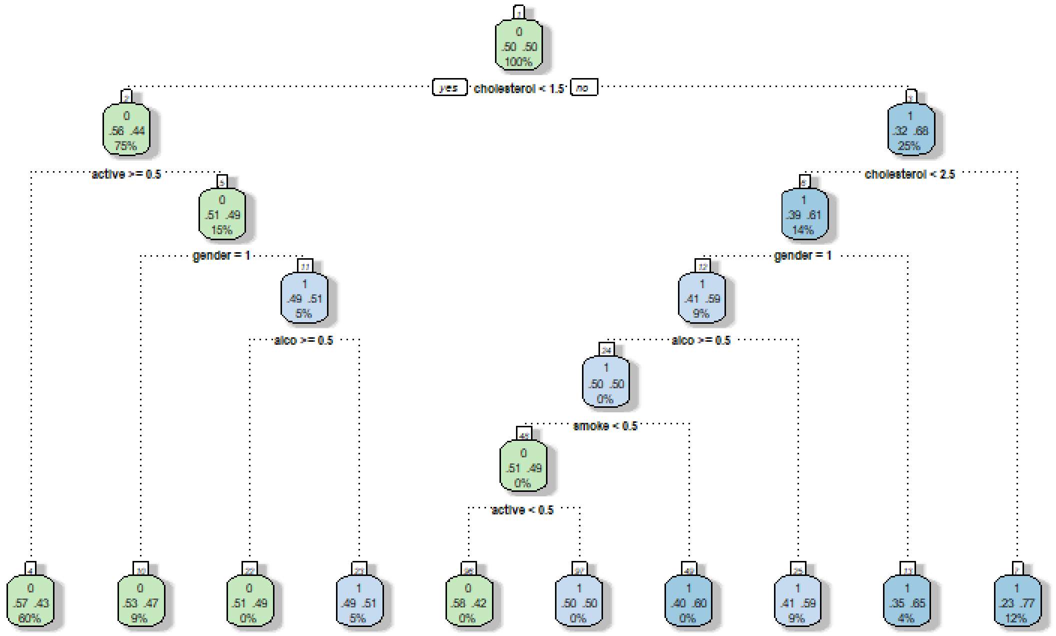
I. Introduction

The dataset selected contains 70000 records on patient's data in relation to Cardiovascular Disease. There are twelve columns, of which Patient ID has been removed for the purpose of my analysis as it not a descriptive variable. Based on the data contained within this dataset, I felt it would be interesting to create a model for the purpose of predicting the likelihood someone has or does not have cardiovascular disease (1 – yes, 0 – no). To this end, I have selected the following columns for my predictive model and with computational constraints in mind:

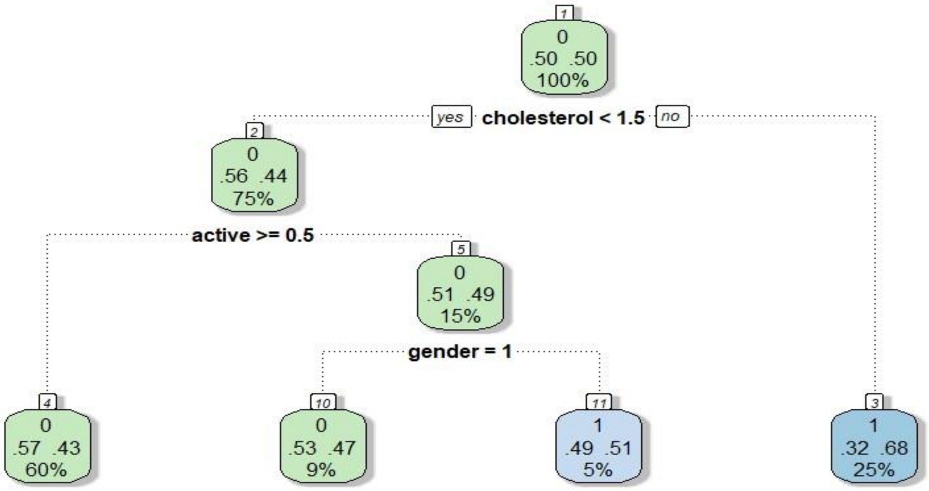
- Cardio
- Active
- Cholesterol
- Smoke
- Alcohol

II. Decision Tree

A 70:30 split was chosen for the purpose of creating training and testing data. In doing this, I accepted the trade off in greater accuracy and a reduced positive rate than if I were to have chosen a 80:20 split.



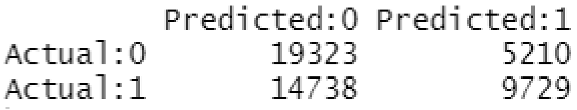
The decision tree above was further refined by identifying and selecting the best complexity parameter (after which reductions in relative error are non-existent or miniscule). The resulting decision tree does not contain the smoking ('smoke') variable or alcohol ('alco') variable, that is the probability of having cardiovascular disease remains roughly the same with these variables removed:



It is interesting to note that in those who have a cholesterol level of less than 1.5 (almost normal) roughly 25% have cardiovascular disease. Of those who have a cholesterol level of greater than 1.5 (above normal) and are active, 60% do not have cardiovascular disease. Of those who have a high level of cholesterol, are inactive and are male, 5% have cardiovascular disease.

III. Evaluation of model

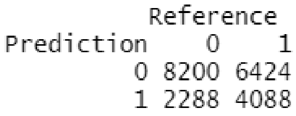
In order to evaluate the performance of this classification model, a confusion matrix was created. When this confusion matrix was used to predict cardiovascular disease on the training data, the following results were achieved:



```
         Predicted:0 Predicted:1
Actual:0       19323        5210
Actual:1       14738        9729
```

As can be seen, the total number of accurate results were 29,052 with 5210 false positives and 14738 false negatives resulting in an accuracy rate of 0.592.

The confusion matrix was subsequently applied to our testing data and the following results were achieved:



```
          Reference
Prediction    0    1
         0 8200 6424
         1 2288 4088
```

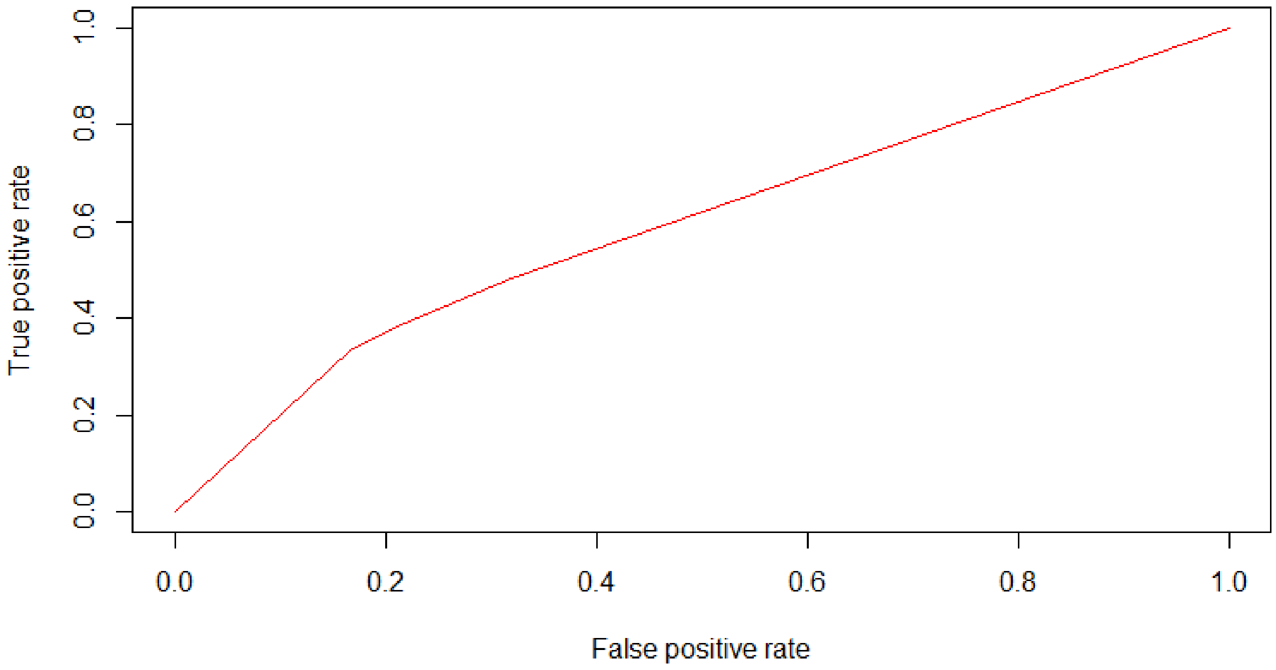
The accuracy of the model decreased to 0.5851 when used on the testing data. The Kappa coefficient decreased from 0.1854 (slight agreement) in the training model to 0.1707 in the testing model. Furthermore, both the Sensitivity and Specificity of the model decreased from 0.7876 and 0.3976 in the training model to 0.7818 and 0.3889 in the testing model respectively. With such a low Kappa coefficient, it would be foolhardy to accept the model in the absence of further, more robust testing.

Given the degree of accurateness of this model, the following conclusions can be reached:

- More data may result in greater accurateness
- A greater selection of variables may increase the completeness and therefore accuracy of our model

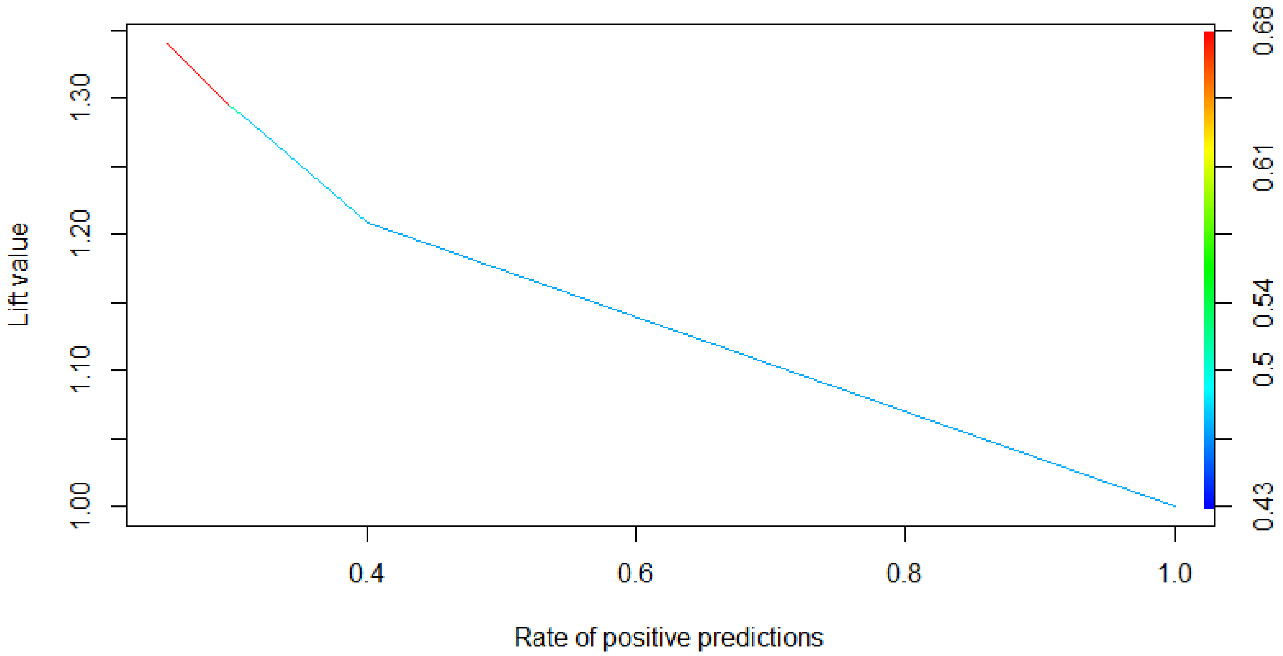
Given the computational limitations experienced, I was unable to confirm the first of these two observations. Subsequent tests on incorporating 1, 2 and 3 more variables decreased the accuracy of the model as opposed to increasing its accuracy.

A Receiver Operating Characteristic (ROC) curve was subsequently developed to illustrate the relationship between false positives (specificity) and true positives (sensitivity) in this model:



As can be seen as the true positive rate increases, so too does the false positive rate. If we were to draw a straight line from (0,0) to (1.1), we would see that the area under the curve is less than the area over the curve.

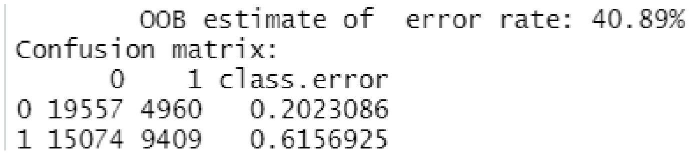
A ‘Lift Chart’ was generated to graphically represent the effectiveness of the predictive model (calculated as the ratio between the results obtained with and without the predictive model):



Finally, the Kolmogorov-Smirnov test was conducted to calculate the maximum different between the cumulative true positive and cumulative false positive rate, generating a score of 0.1763284. In general, the higher this score (to 1) the more accurate the model. Given the low KS value, it can be deduced the model is inaccurate.

IV. Improving the Accuracy of the Model

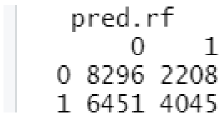
In order to improve the accuracy of the model, the random forest technique was used with the following results obtained when calling the random forest on our selected data:



```
        OOB estimate of  error rate: 40.89%
Confusion matrix:
      0    1 class.error
0 19557 4960   0.2023086
1 15074 9409   0.6156925
```

The out-of-bag error was high at 40.89%, meaning 40.89% of results were inaccurately predicted, very similar to the initial scores when evaluating the model used as in part III. The number of false negatives increased, while the number of false positives decreased. In total there were 28,966 accurate results.

The following results were achieved for the test data:



```
  pred.rf
       0    1
0 8296 2208
1 6451 4045
```

It is interesting to note that the number of false negatives increased substantially, with the number of false positives decreasing in a similar fashion. In spite of this, the accuracy of the model was very similar to that of our pre-random forest results with an accuracy rate of 0.5877 (up from 0.5851). The Kappa value also increased to 0.1752 (from 0.1707) with the Sensitivity rate increasing to 0.7898 (from 0.7818) and Specificity rate decreasing to 0.3854 from 0.3889 which makes sense given where one increases, the other decreases (the so called ‘trade-off’ referred to earlier).

V. Conclusion

In conclusion, the model developed and analyzed is not very accurate, and were it to be applied in real life – roughly 40% of people who display the characteristics outlined in the variables selected would be falsely identified as having cardiovascular disease. This is of course unacceptable in a medical context where accuracy is paramount. Greater data could perhaps result in a more accurate model as touched on earlier however this would require greater computational power. Further models could be investigated, perhaps including, or removing certain variable(s) however this is beyond the scope of this evaluation.

VI. Appendix

https://www.kaggle.com/sulianova/cardiovascular-disease-dataset - Source for dataset