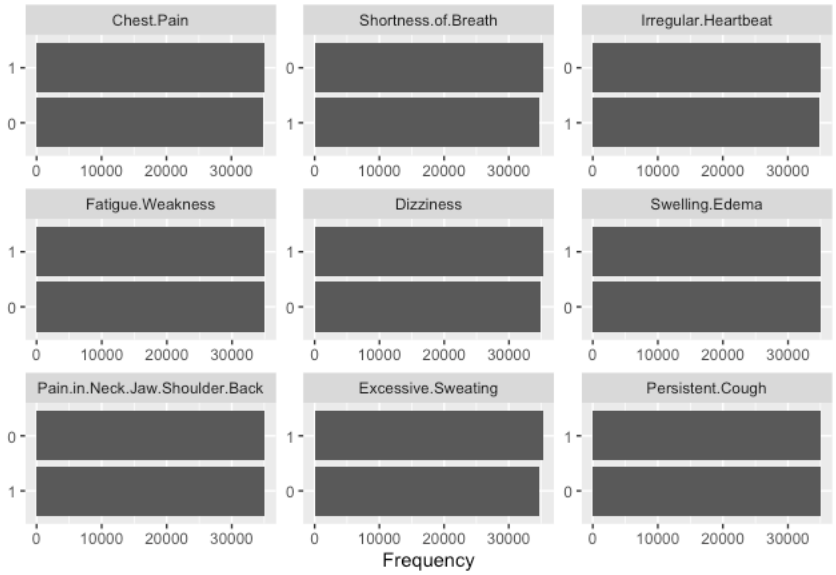
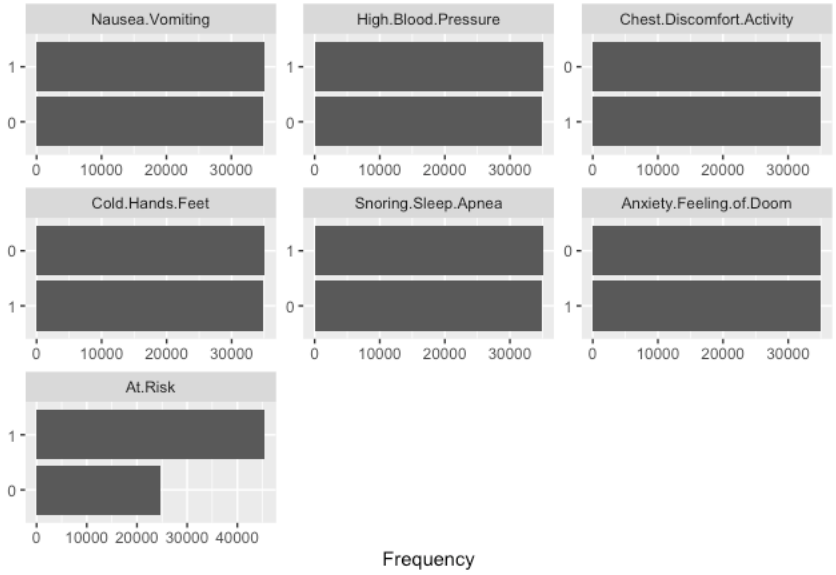
Supplementary Appendix

Figure 1: Total number of 1's and 0's for each categorical variable



Page 1



Page 2

Figure 2: Distribution of Stroke.Risk.Percentage and Age

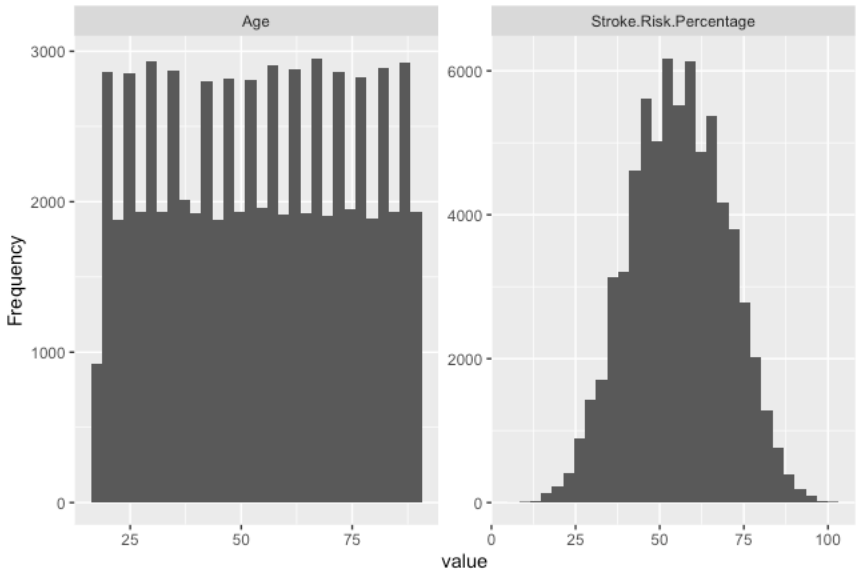


Figure 3: Example of how K-folds cross validation works (in this example, k = 5)

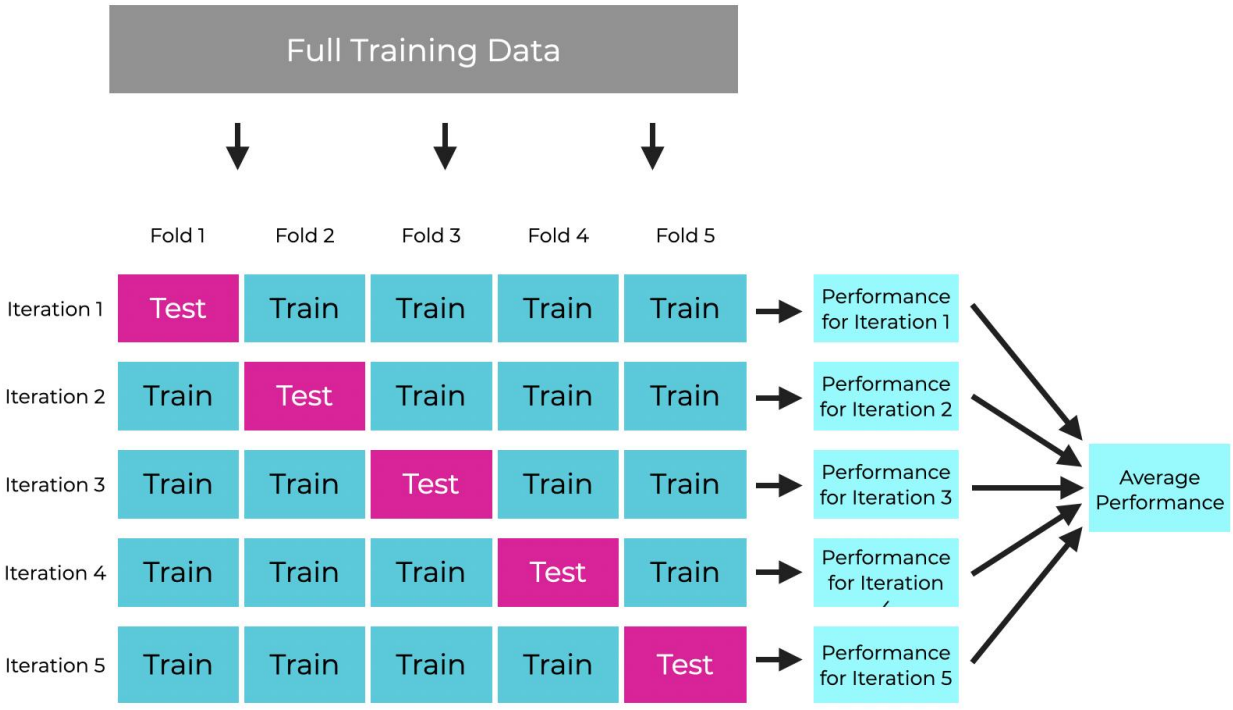
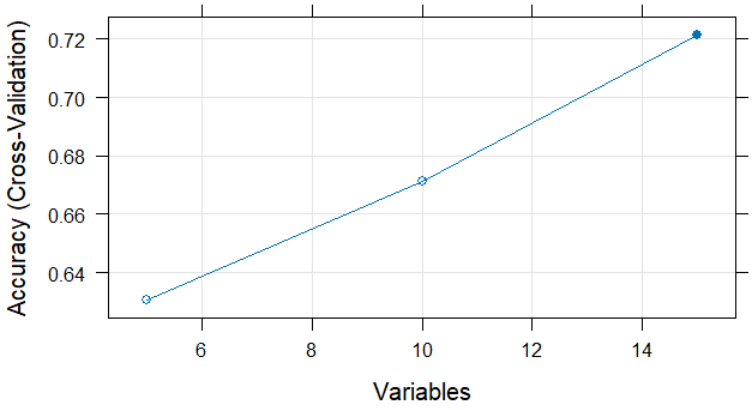


Figure 4: RFE results for logistic regression




```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables Accuracy  Kappa AccuracySD KappaSD Selected
         5   0.6307 0.2616   0.005410 0.01082
        10   0.6711 0.3419   0.005431 0.01087
        15   0.7212 0.4427   0.007563 0.01513        *

The top 5 variables (out of 15):
   Cold.Hands.Feet, Anxiety.Feeling.of.Doom, Snoring.Sleep.Apnea, High.Blood.Pressure, Excessive.Sweating
```

Figure 5: LASSO and Ridge Regression Feature importance

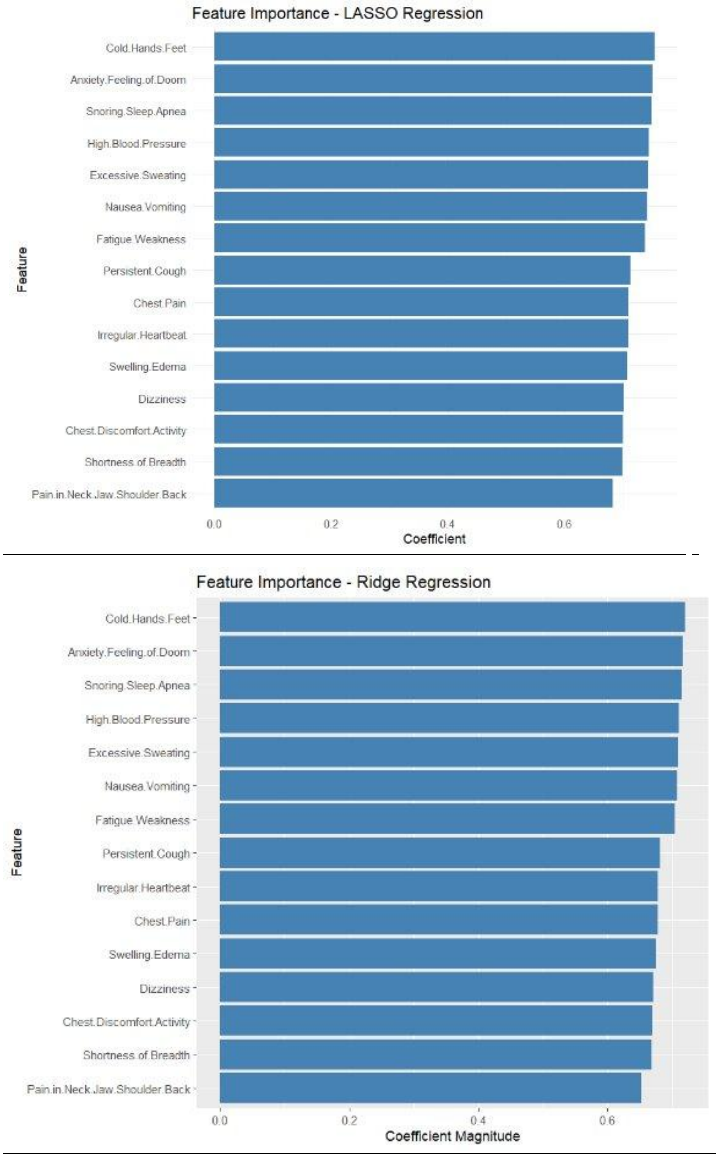


Figure 6: Elastic Net Feature importance

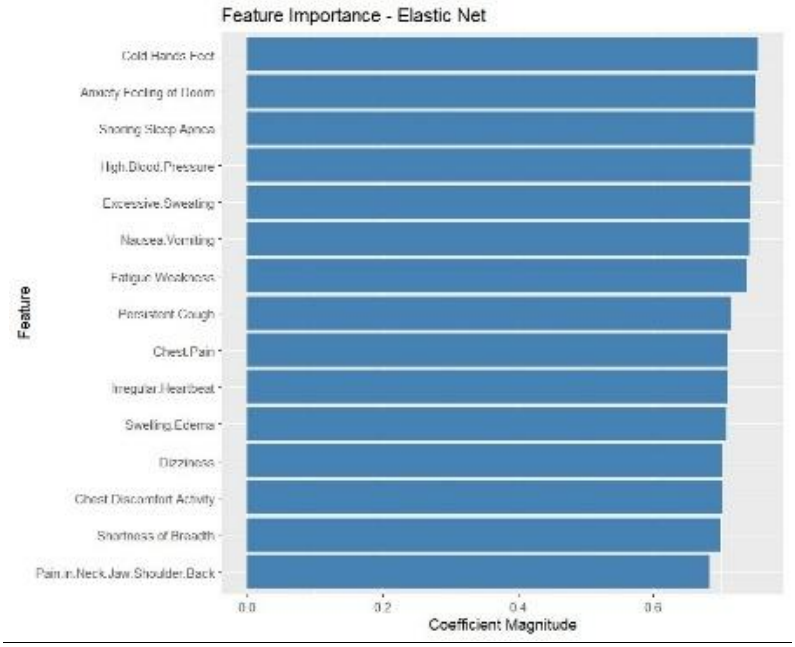


Figure 7: Example of a decision tree created on the dataset

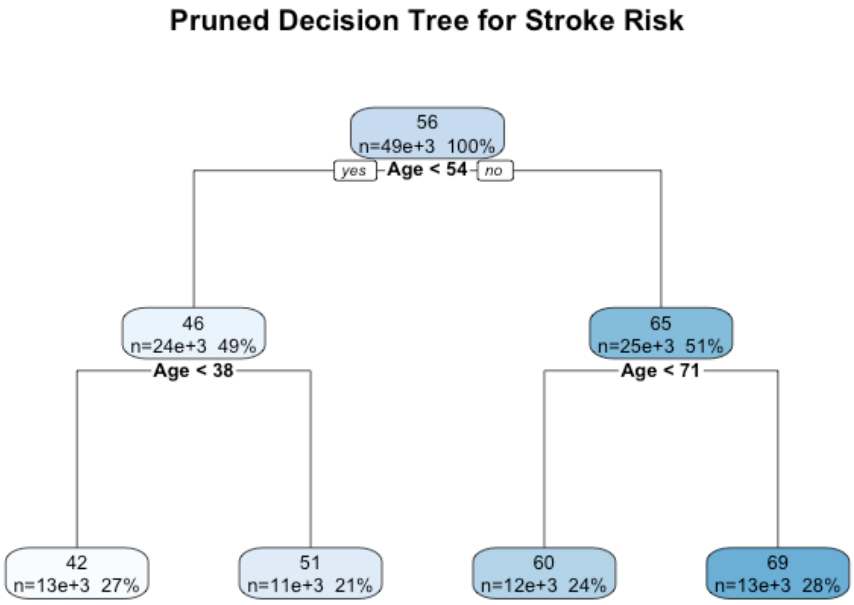


Figure 8: Decision tree after Age was removed

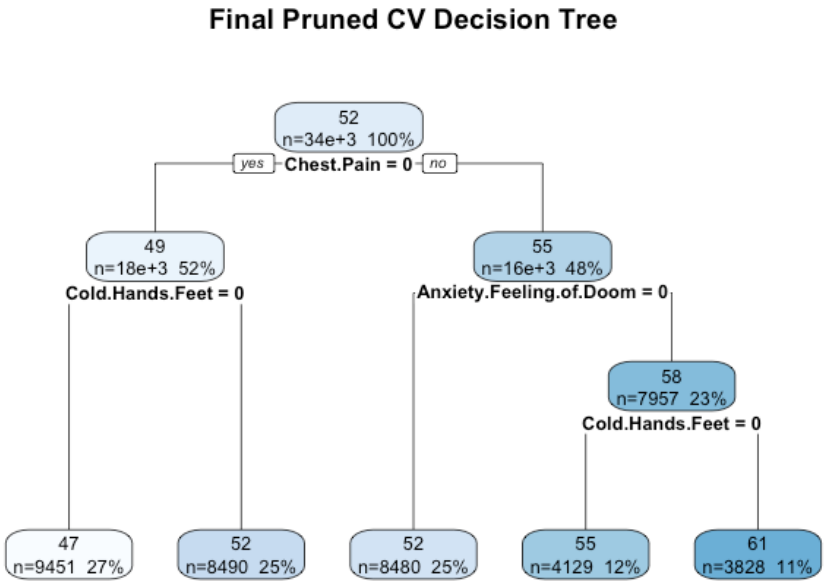


Figure 9: Gradient Boosting Feature Importances - (graphs below are in order from left to right and from top to bottom the full model without Age, for the top 8 predictors, and for the top 9 predictors)