Stroke Prediction

Shraddha P Jain

20-PBD-002

Objective

- According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.
- The objective of this modelling is to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status, etc.

Dataset

1) id: unique identifier

2) gender: "Male", "Female" or "Other"

3) age: age of the patient

4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6) ever_married: "No" or "Yes"

7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"

8) Residence_type: "Rural" or "Urban"

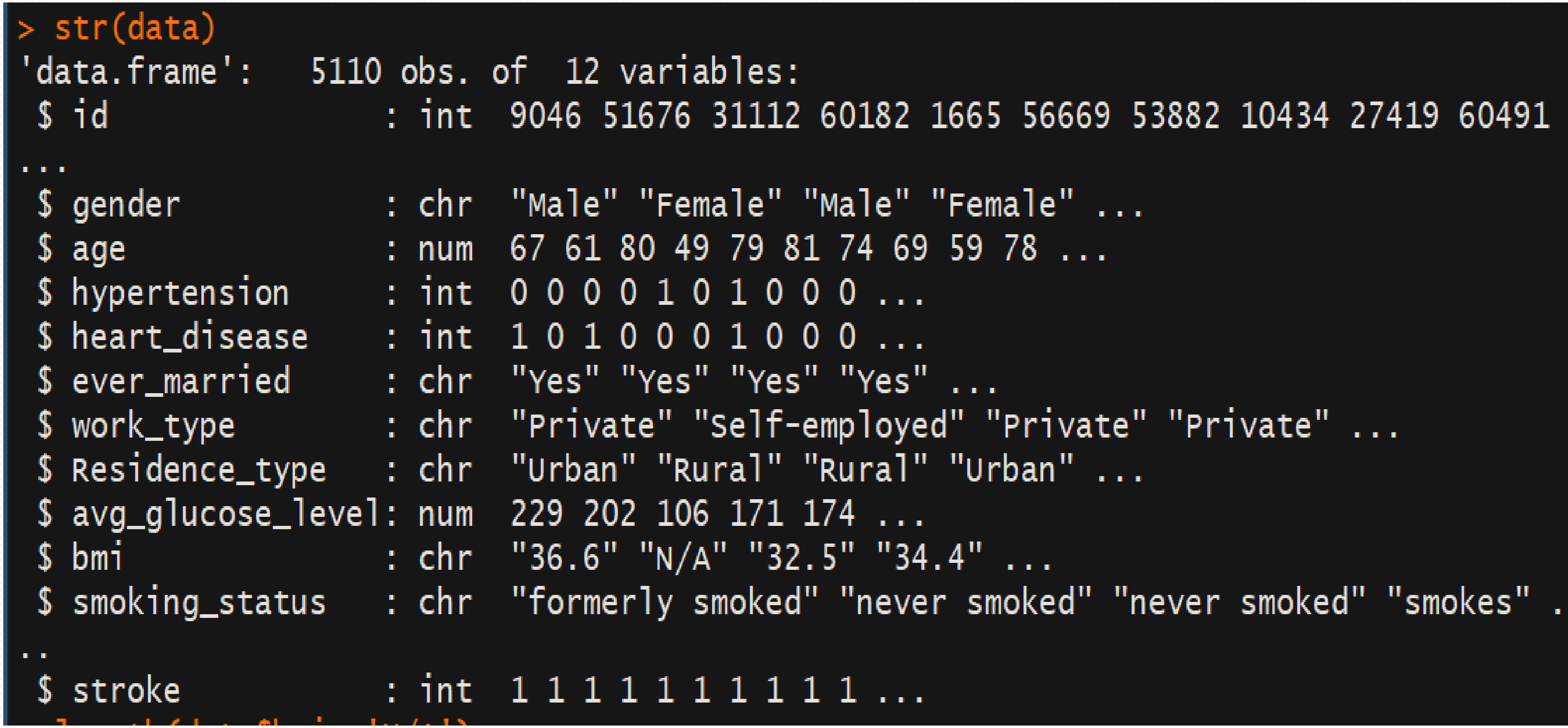
9) avg_glucose_level: average glucose level in blood

10) bmi: body mass index

11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*

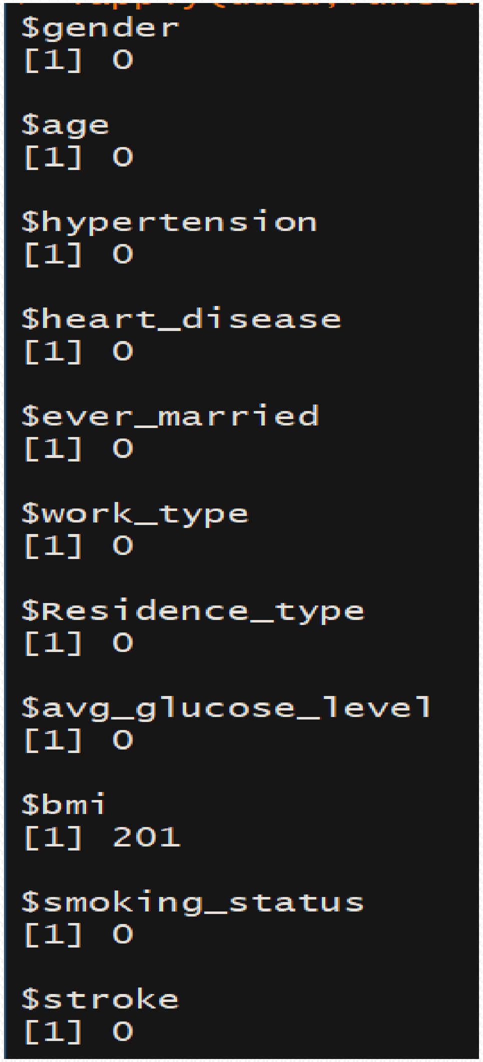
12) stroke: 1 if the patient had a stroke or 0 if not

Data Preprocessing

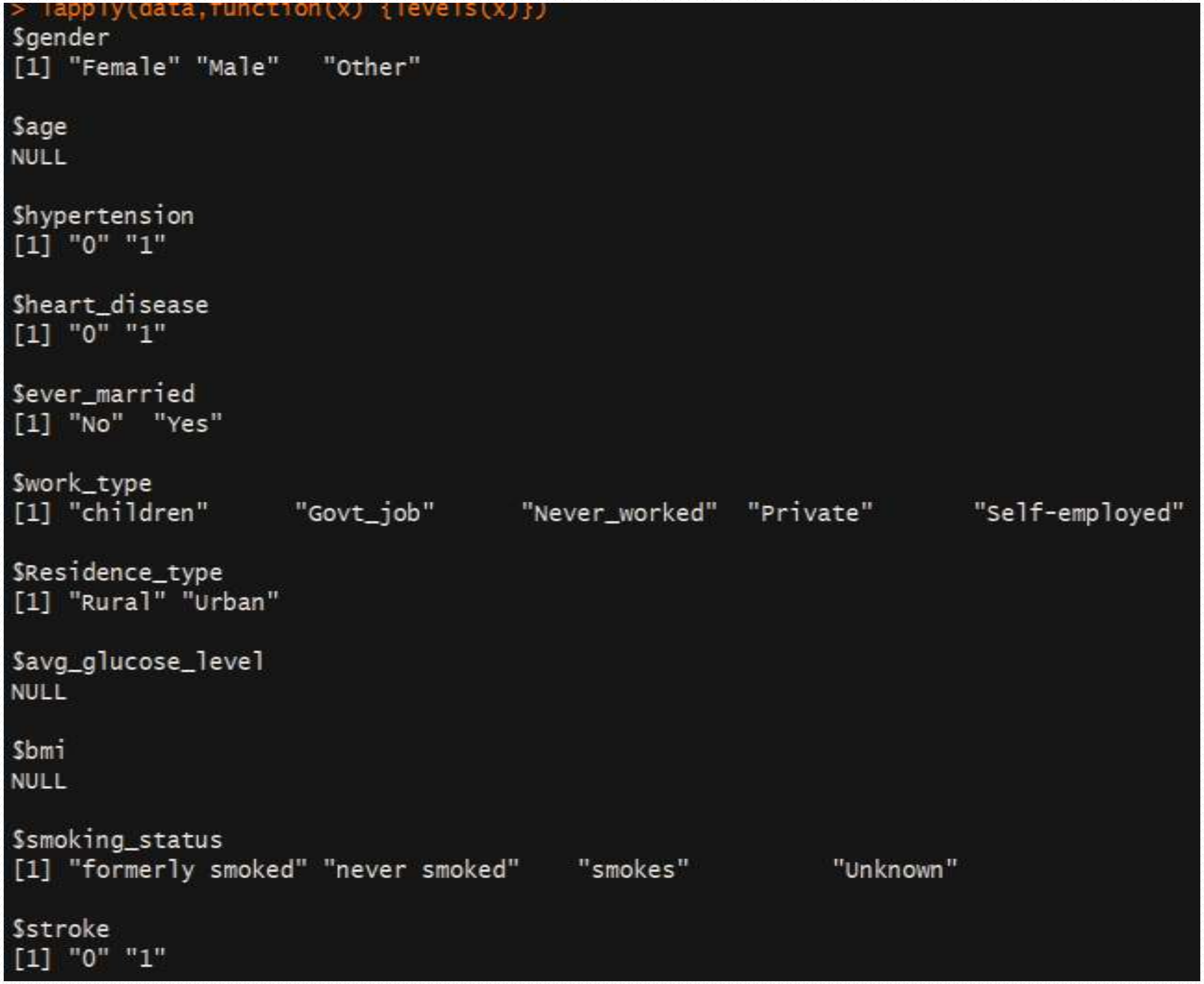


```
> str(data)
'data.frame':	5110 obs. of  12 variables:
 $ id               : int  9046 51676 31112 60182 1665 56669 53882 10434 27419 60491
...
 $ gender           : chr  "Male" "Female" "Male" "Female" ...
 $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
 $ hypertension     : int  0 0 0 0 1 0 1 0 0 0 ...
 $ heart_disease    : int  1 0 1 0 0 0 1 0 0 0 ...
 $ ever_married     : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ work_type        : chr  "Private" "Self-employed" "Private" "Private" ...
 $ Residence_type   : chr  "Urban" "Rural" "Rural" "Urban" ...
 $ avg_glucose_level: num  229 202 106 171 174 ...
 $ bmi              : chr  "36.6" "N/A" "32.5" "34.4" ...
 $ smoking_status   : chr  "formerly smoked" "never smoked" "never smoked" "smokes" .
..
 $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

Checking for NA’s and Levels

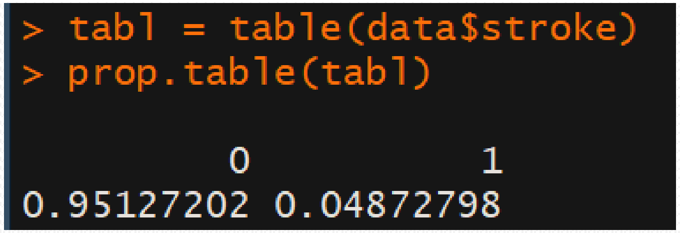


```
$gender
[1] 0

$age
[1] 0

$hypertension
[1] 0

$heart_disease
[1] 0

$ever_married
[1] 0

$work_type
[1] 0

$Residence_type
[1] 0

$avg_glucose_level
[1] 0

$bmi
[1] 201

$smoking_status
[1] 0

$stroke
[1] 0
```

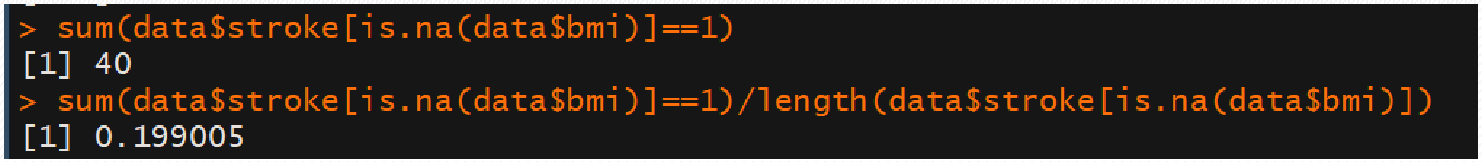


```
> lapply(data,function(x) {levels(x)})
$gender
[1] "Female" "Male"   "Other"

$age
NULL

$hypertension
[1] "0" "1"

$heart_disease
[1] "0" "1"

$ever_married
[1] "No"  "Yes"

$work_type
[1] "children"      "Govt_job"      "Never_worked"  "Private"       "Self-employed"

$Residence_type
[1] "Rural" "Urban"

$avg_glucose_level
NULL

$bmi
NULL

$smoking_status
[1] "formerly smoked" "never smoked"    "smokes"          "Unknown"

$stroke
[1] "0" "1"
```

Dealing with NA’s



```
> tabl = table(data$stroke)
> prop.table(tabl)

         0          1 
0.95127202 0.04872798 
```



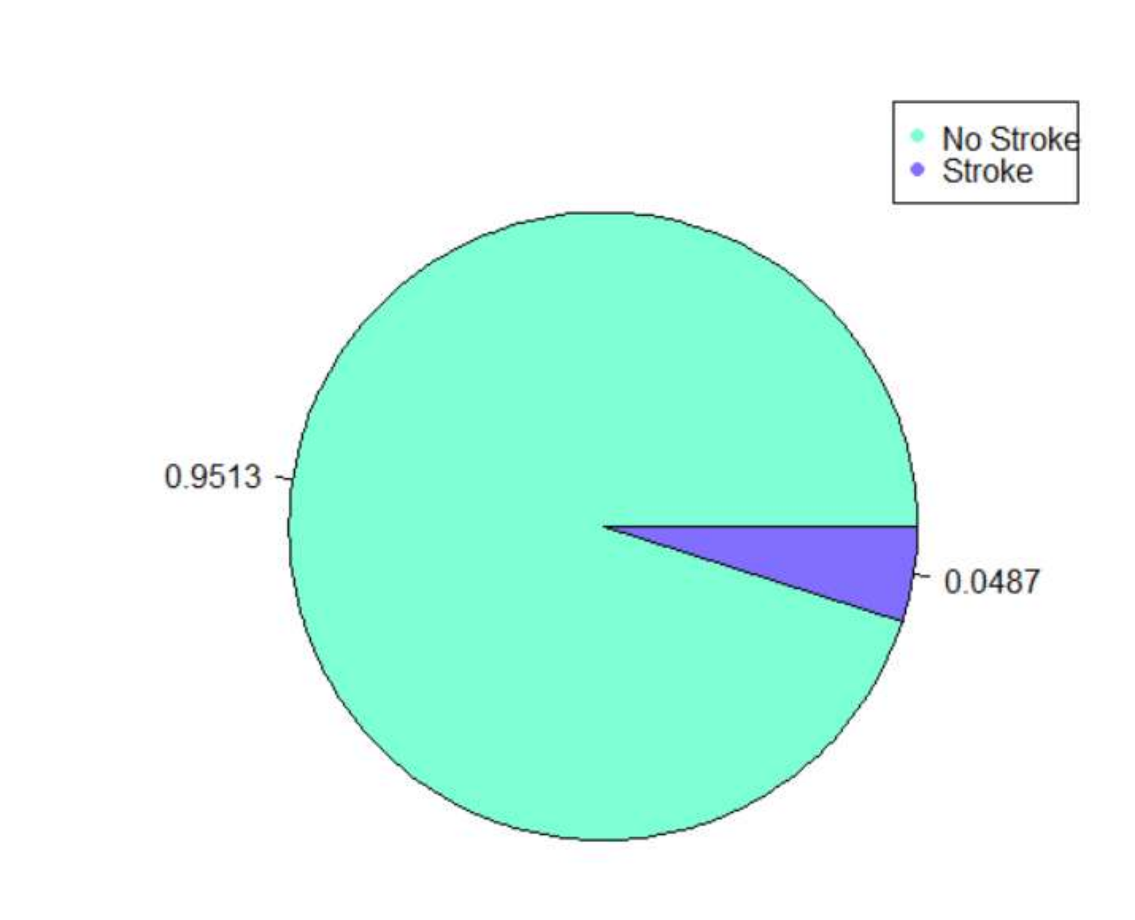
```
> sum(data$stroke[is.na(data$bmi)]==1)
[1] 40
> sum(data$stroke[is.na(data$bmi)]==1)/length(data$stroke[is.na(data$bmi)])
[1] 0.199005
```

- Imputation using MICE

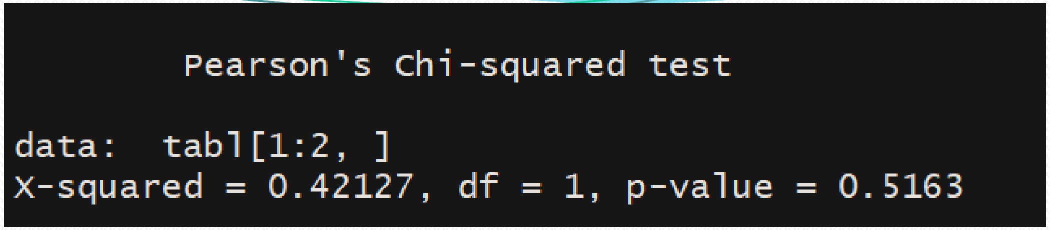
Exploratory data Analysis

- For gender ‘Other’, there is only 1 patient who has not had any attack of stroke.
- For the work_type, due to not-enough samples, we have only 22 patients who have never worked, and all of them did not have an attack of stroke.

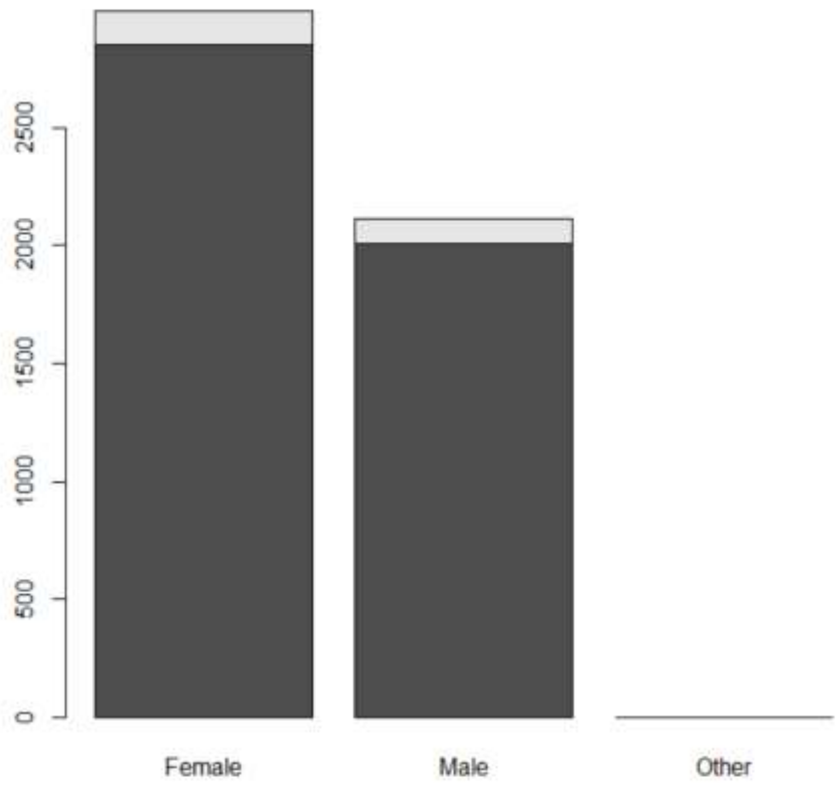
Stroke

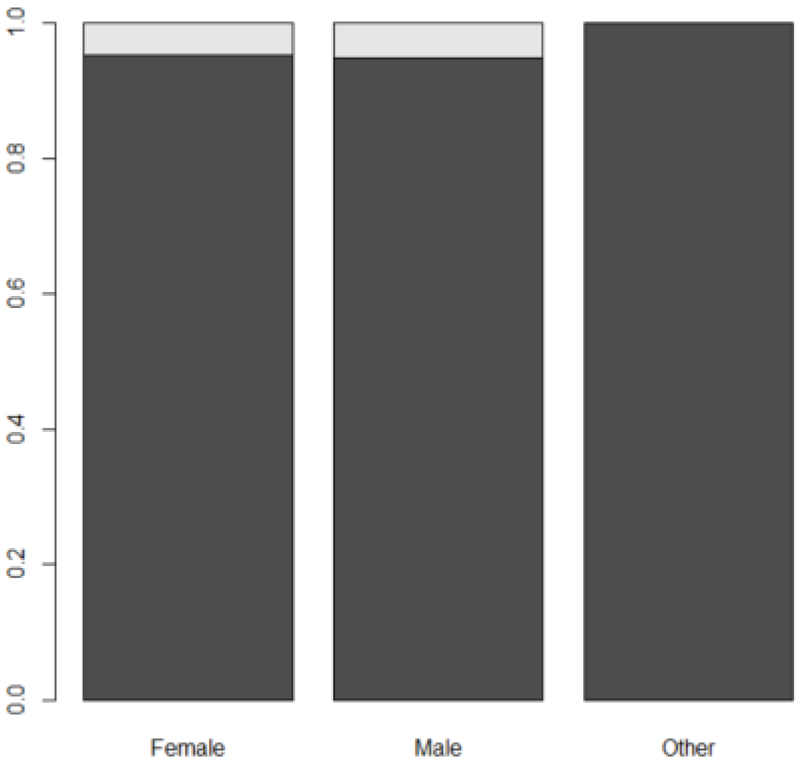


Gender

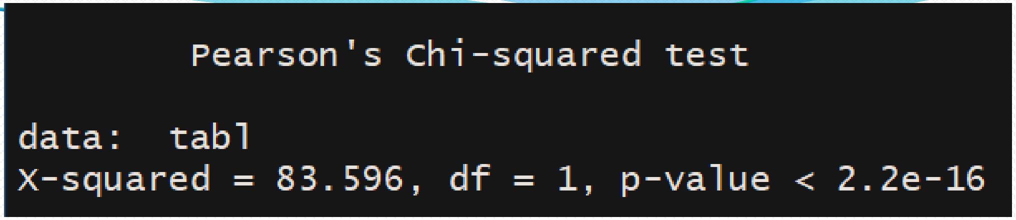


```
        Pearson's Chi-squared test

data:  tabl[1:2, ]
X-squared = 0.42127, df = 1, p-value = 0.5163
```

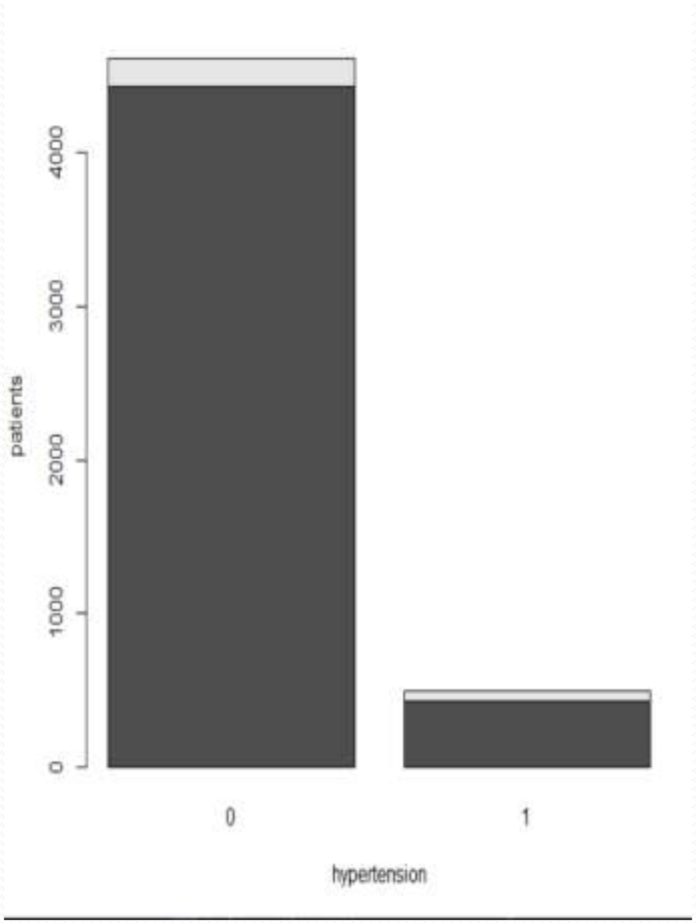


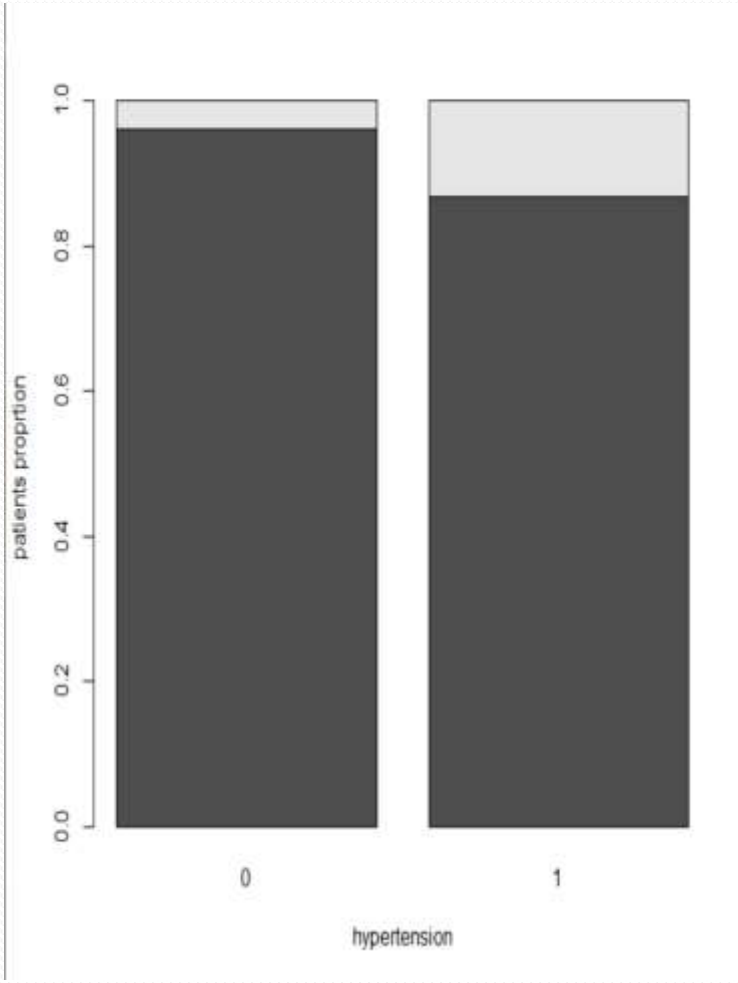


Hypertension

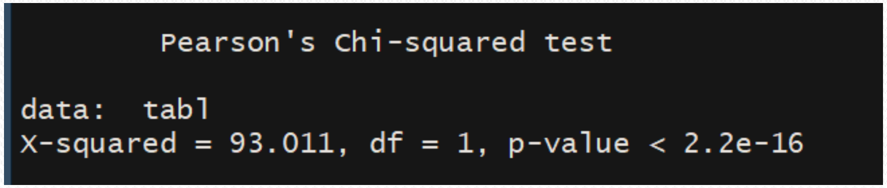


```
        Pearson's Chi-squared test

data:  tabl
X-squared = 83.596, df = 1, p-value < 2.2e-16
```

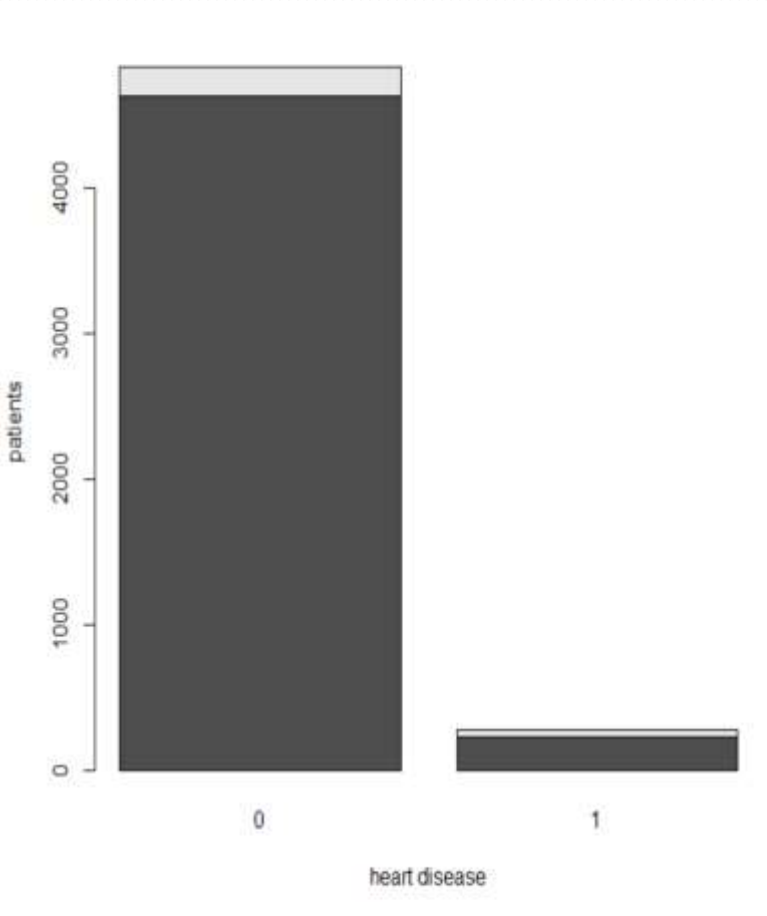


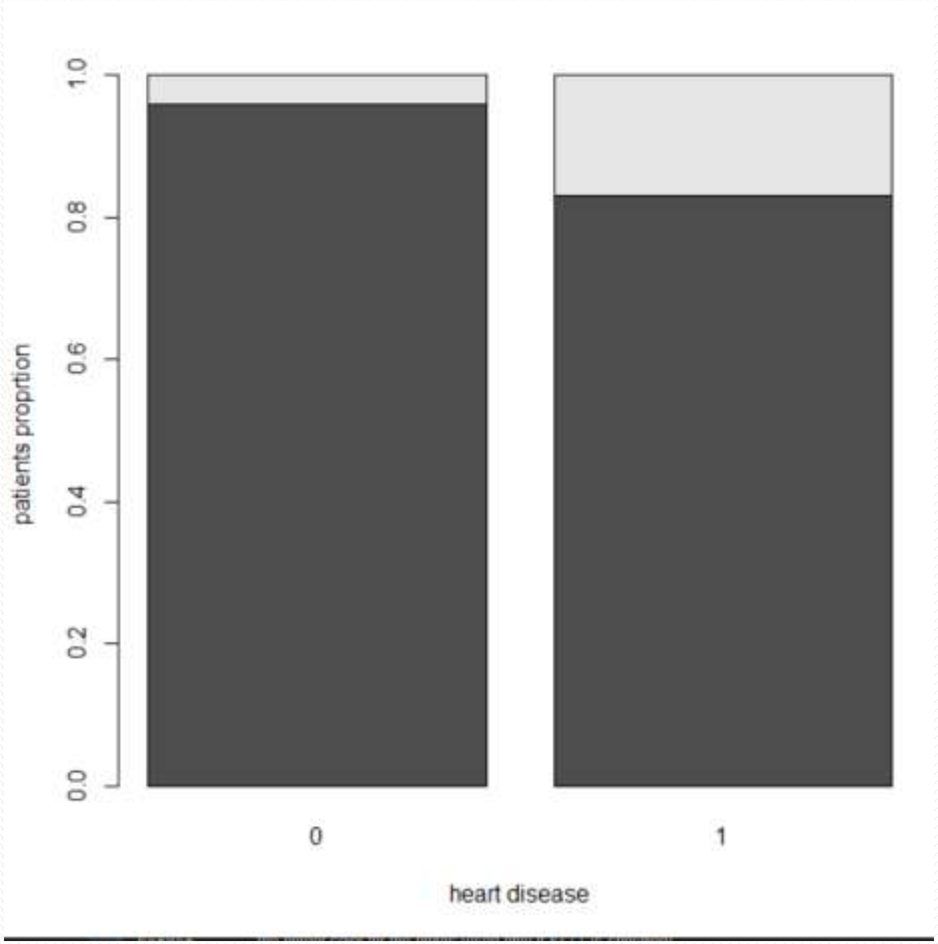


Heart Disease

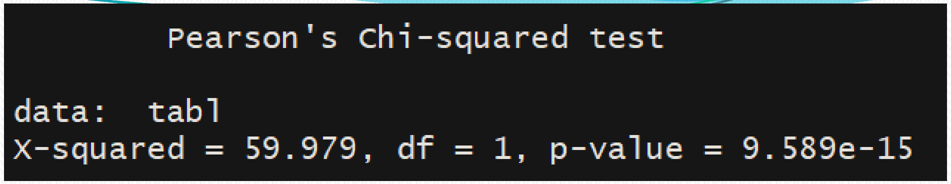


```
        Pearson's Chi-squared test

data:  tabl
X-squared = 93.011, df = 1, p-value < 2.2e-16
```

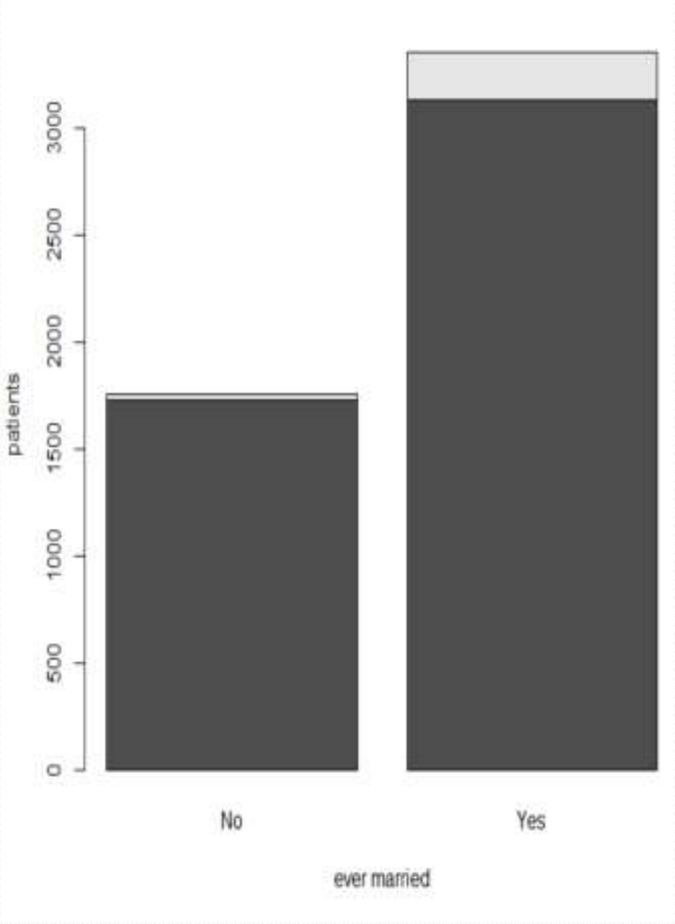


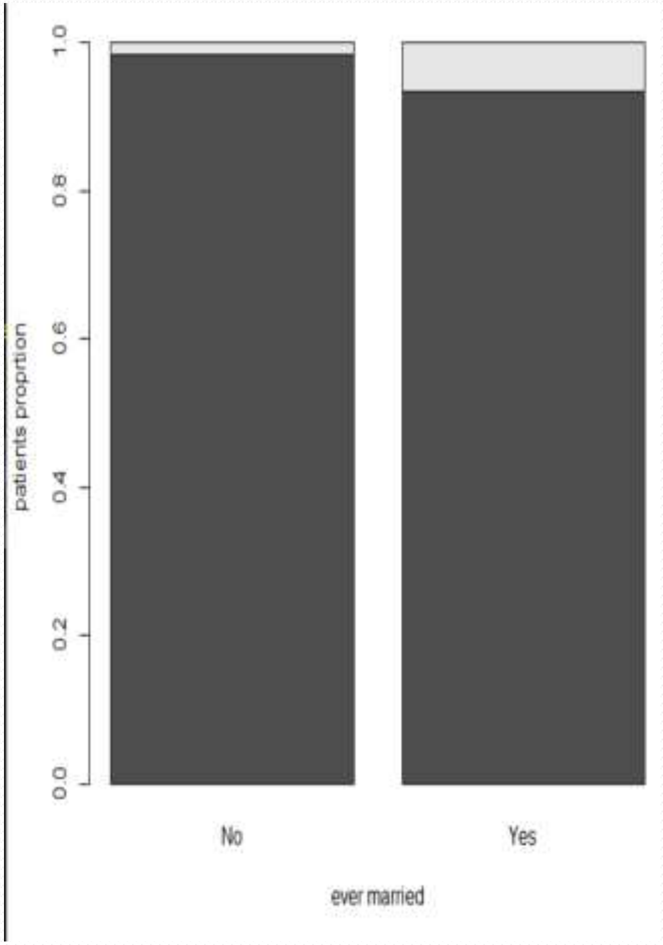


Ever_married

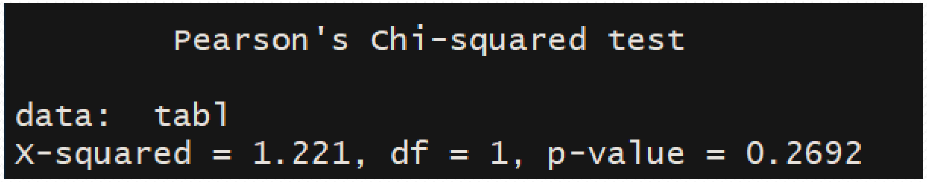


```
        Pearson's Chi-squared test

data:  tabl
X-squared = 59.979, df = 1, p-value = 9.589e-15
```

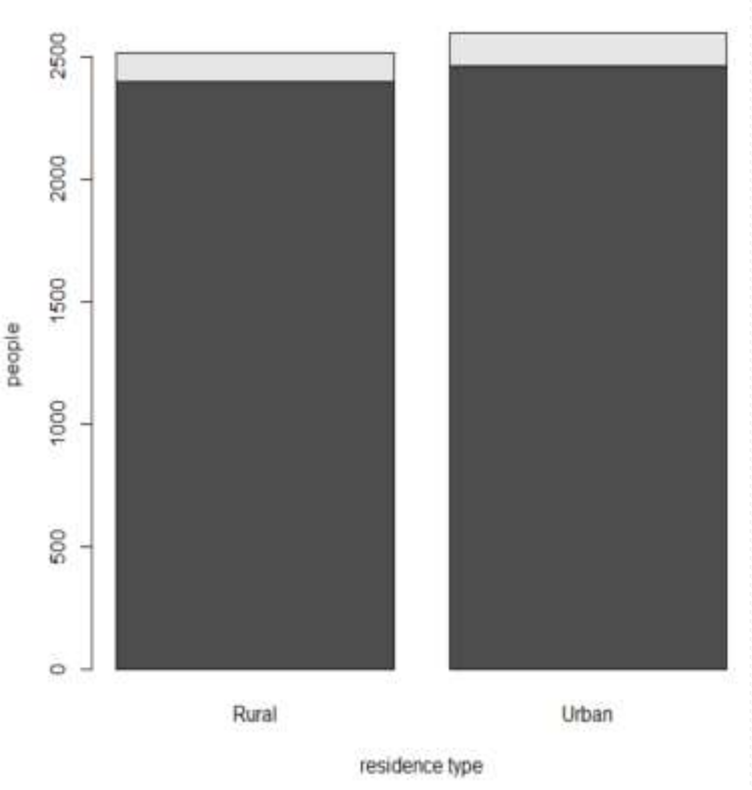


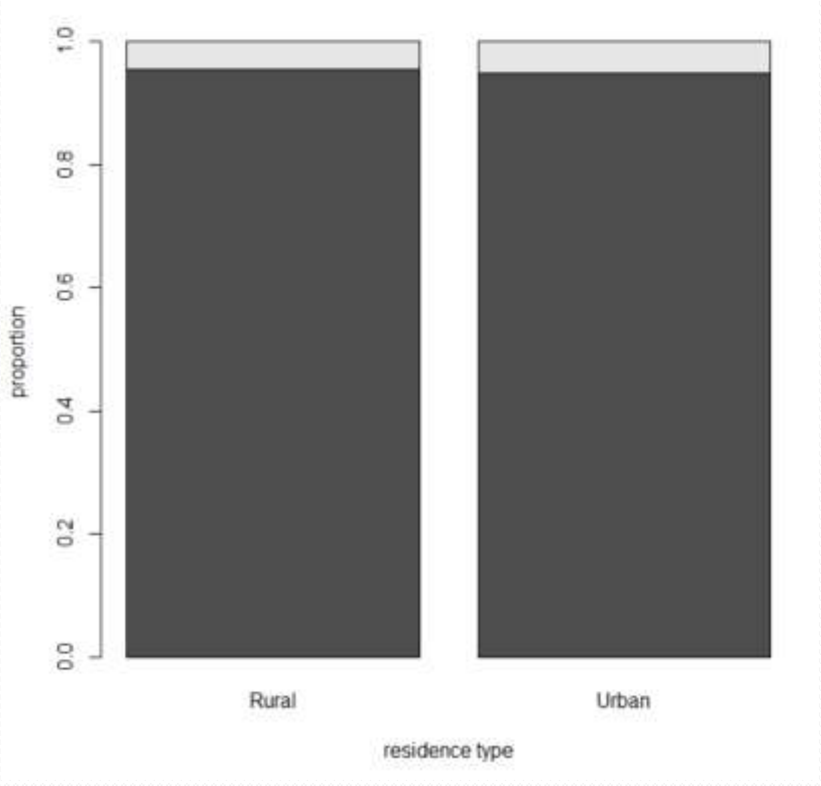


Residence Type

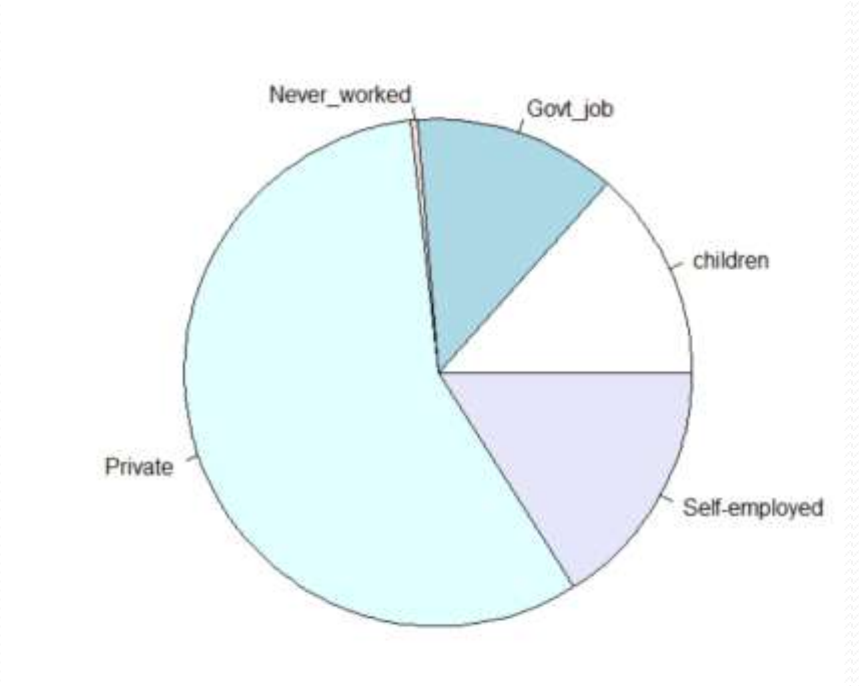


```
        Pearson's Chi-squared test

data:  tabl
X-squared = 1.221, df = 1, p-value = 0.2692
```

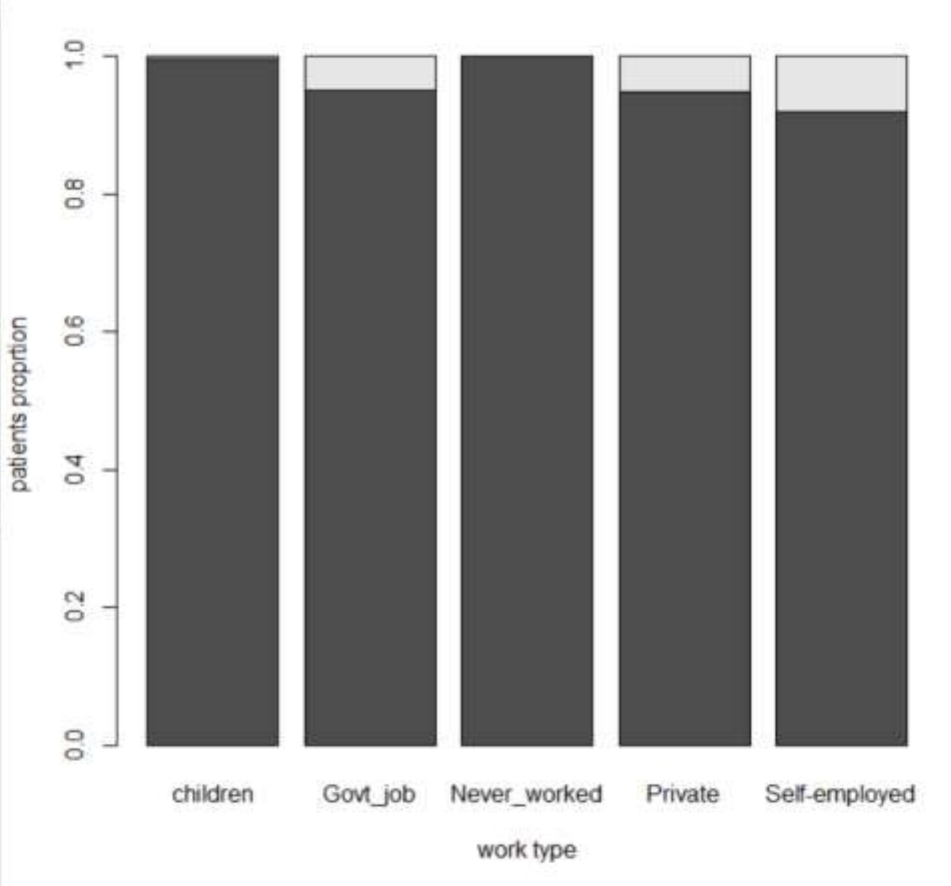




Work Type

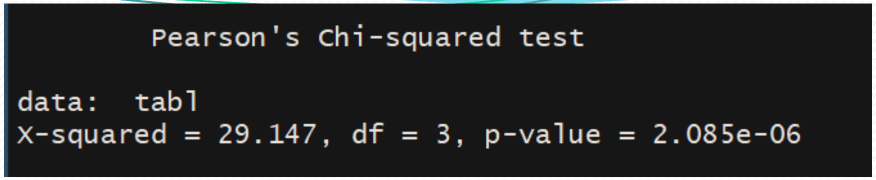


Never_worked
Govt_job
children
Private
Self-employed

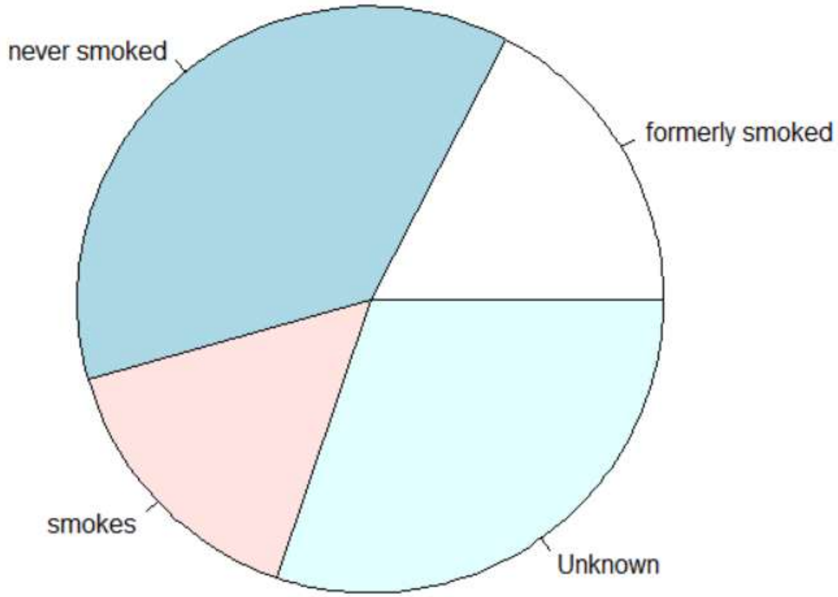


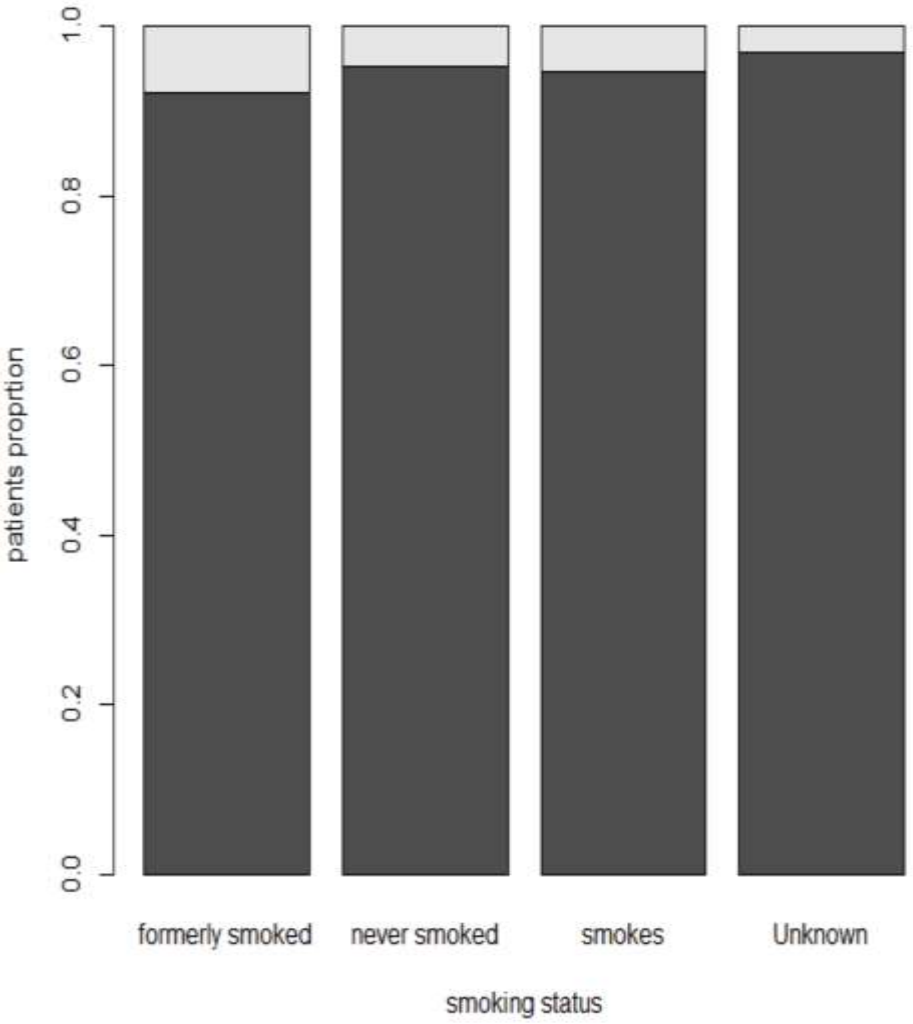
patients proprtion
0.0
0.2
0.4
0.6
0.8
1.0
children
Govt_job
Never_worked
Private
Self-employed
work type

Smoking Status

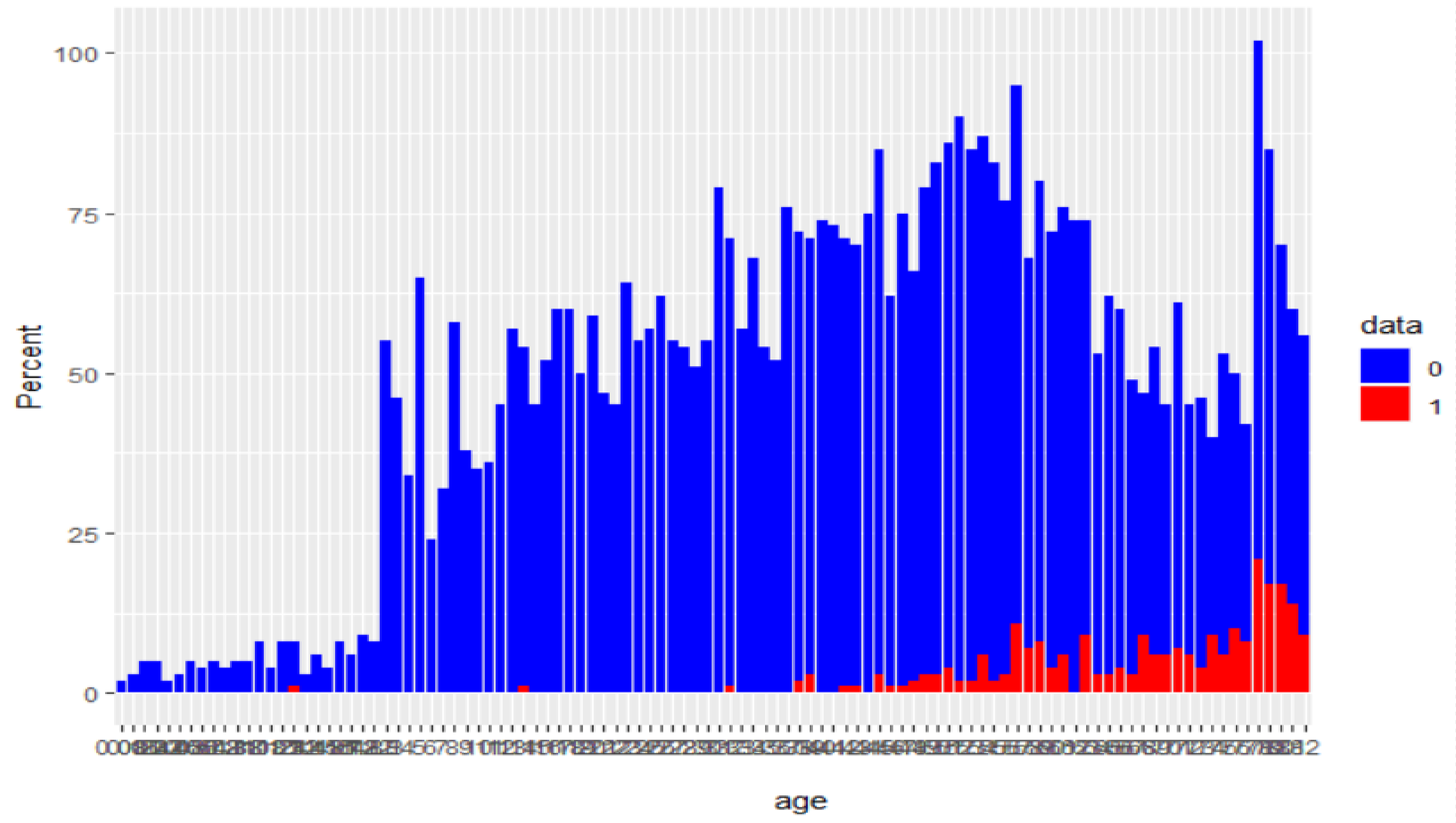


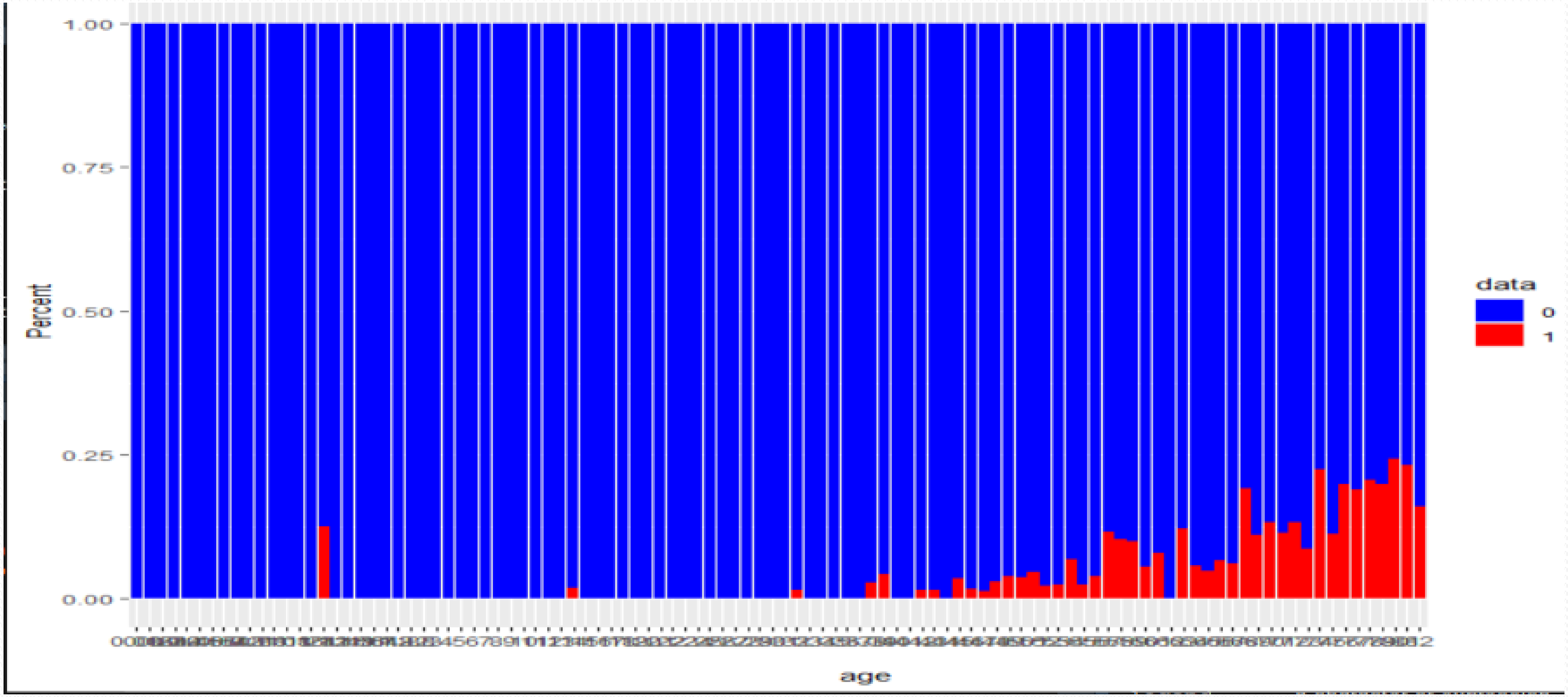
```
        Pearson's Chi-squared test

data:  tabl
X-squared = 29.147, df = 3, p-value = 2.085e-06
```

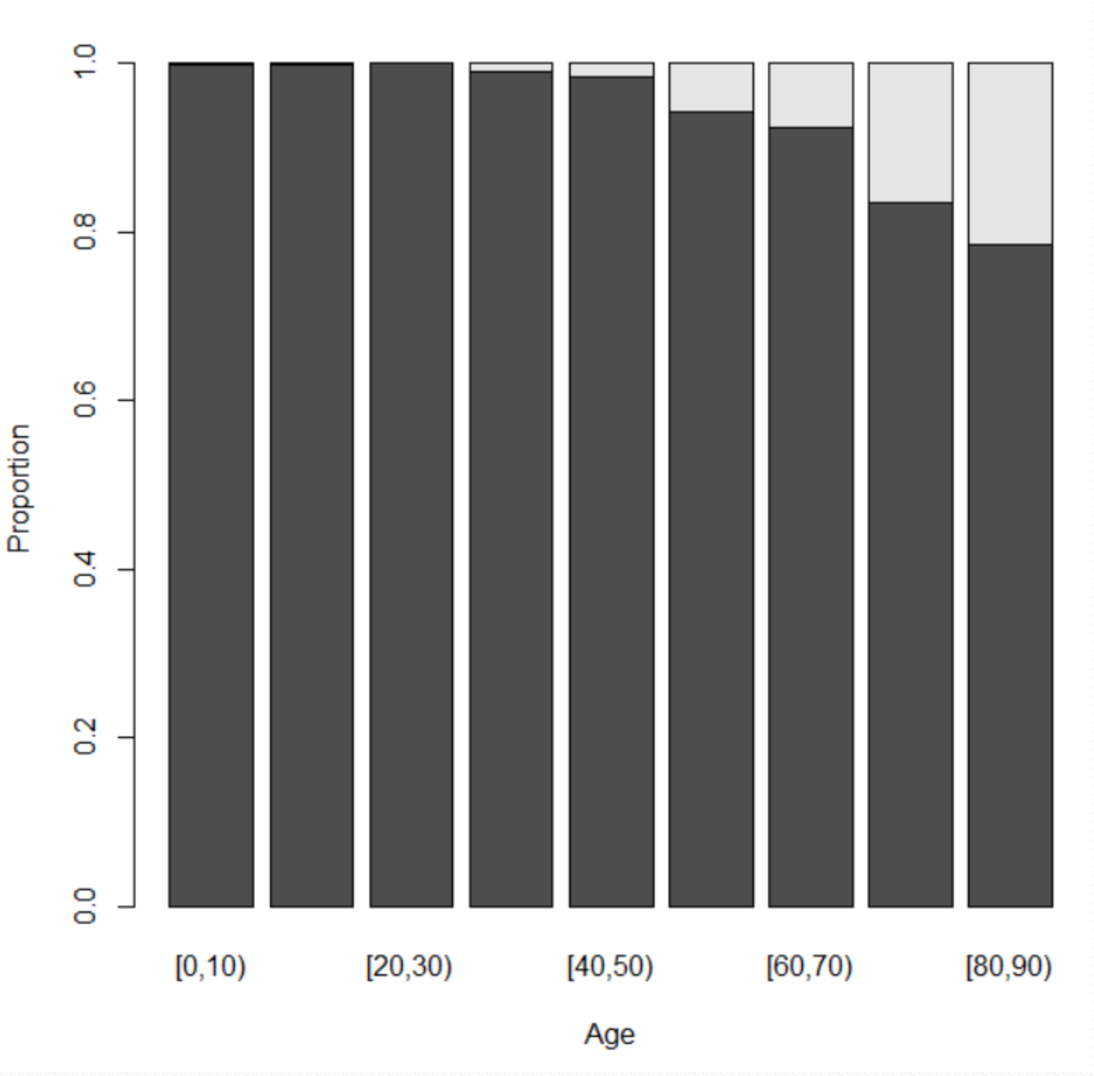




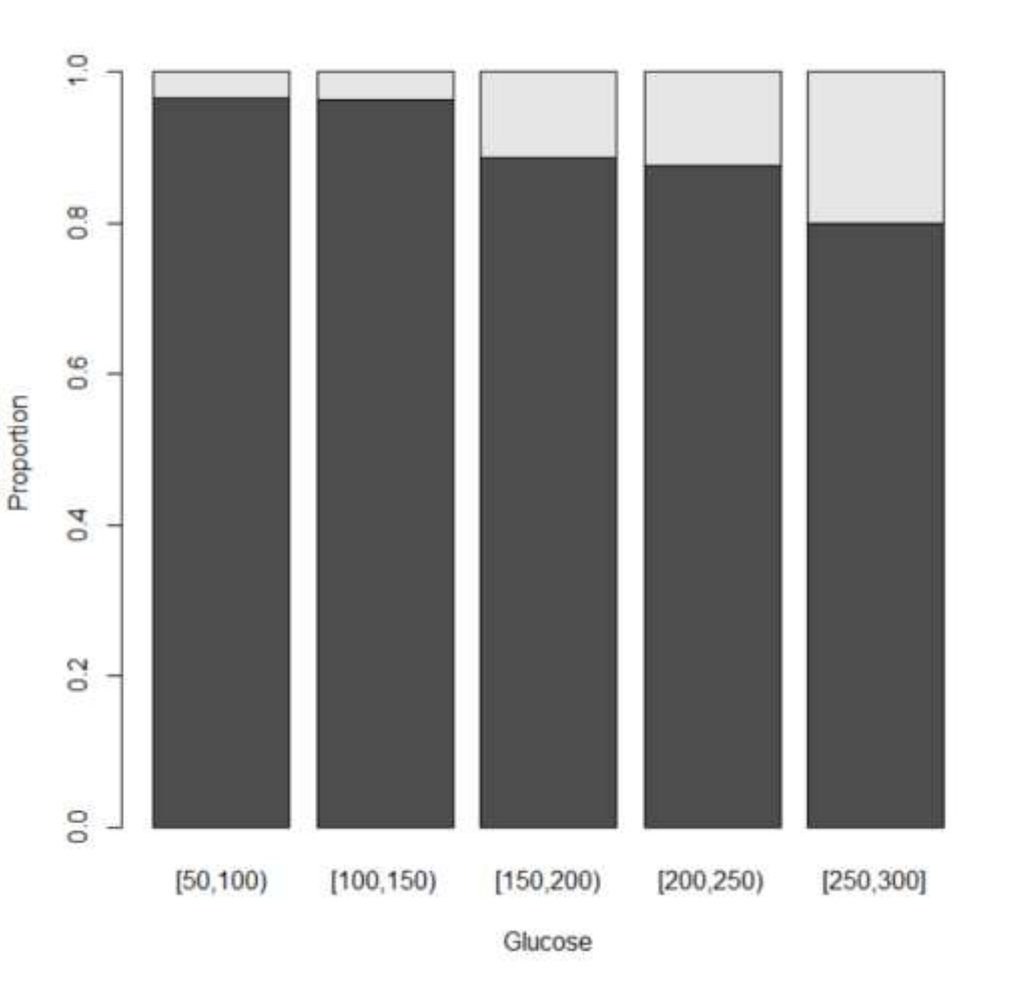
Age



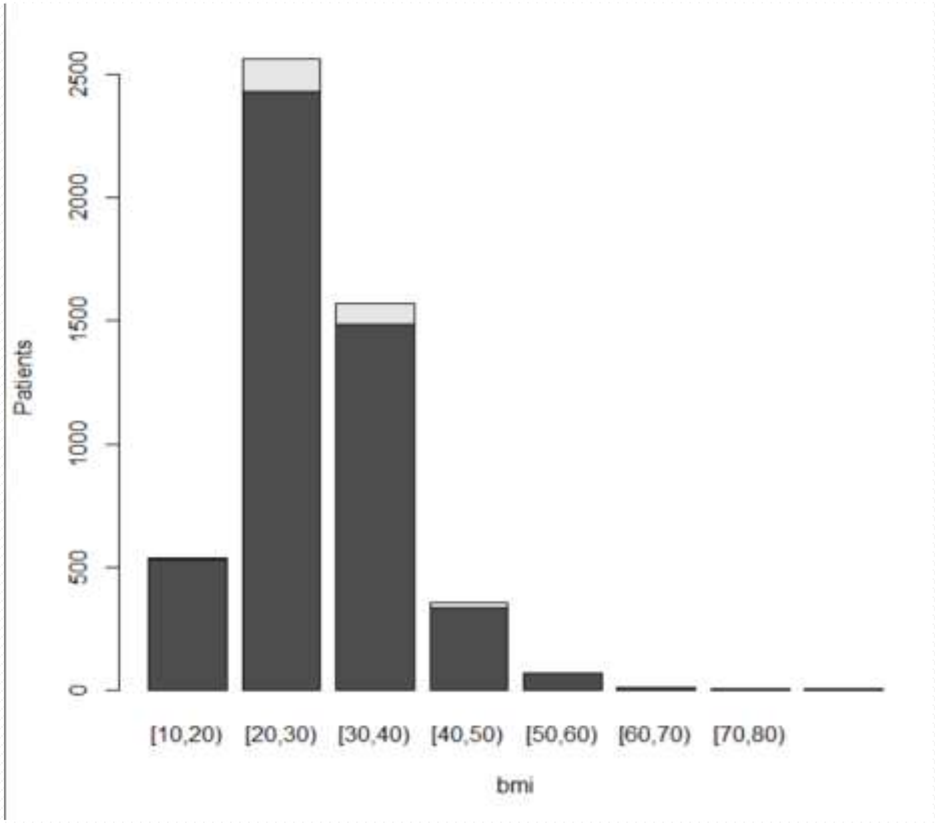

1.00
0.75
0.50
0.25
0.00
Percent
age
data
0
1

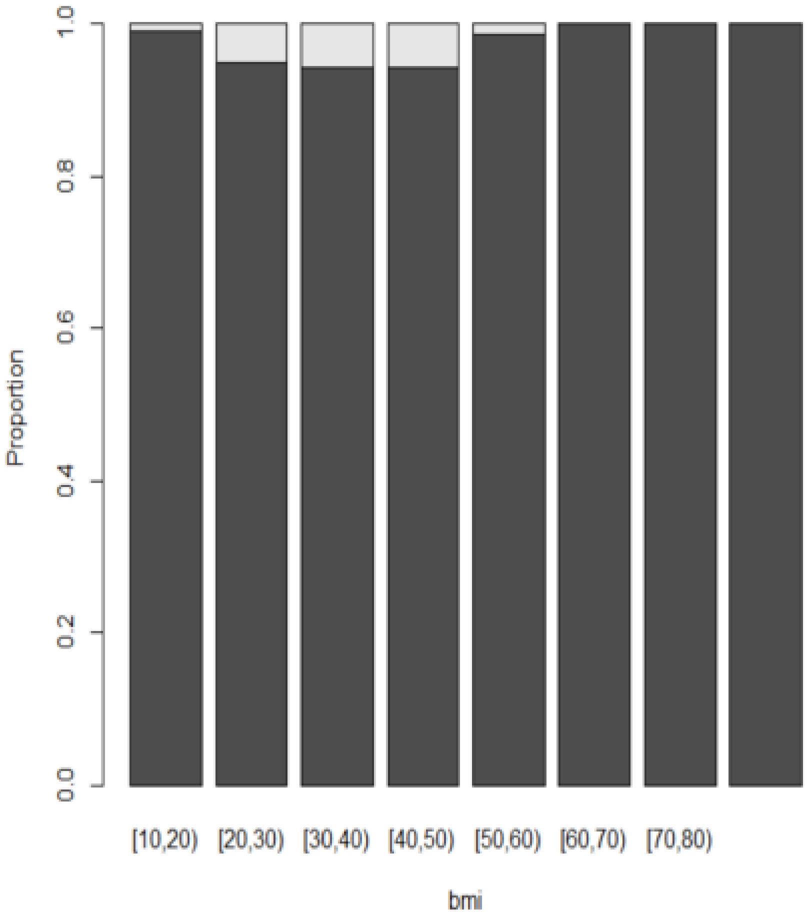

1.0
0.8
0.6
0.4
0.2
0.0
Proportion
[0,10)
[20,30)
[40,50)
[60,70)
[80,90)
Age

Glucose levels binned



BMI binned


Patients
0
500
1000
1500
2000
2500
[10,20)
[20,30)
[30,40)
[40,50)
[50,60)
[60,70)
[70,80)
bmi

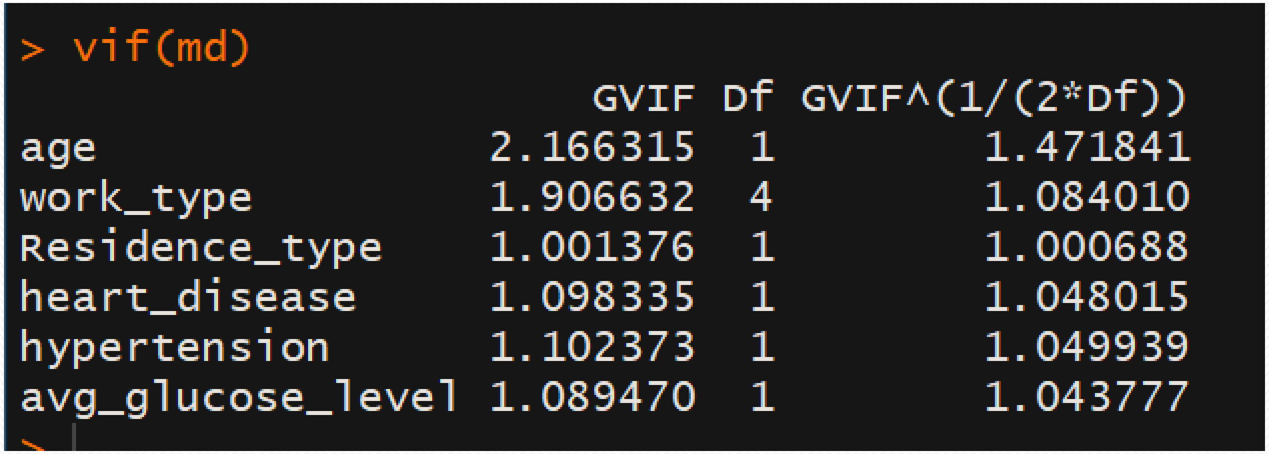

Proportion
0.0
0.2
0.4
0.6
0.8
1.0
[10,20)
[20,30)
[30,40)
[40,50)
[50,60)
[60,70)
[70,80)
bmi

Conclusion

- Hypertension, heartdisease, and ever_married, worktype, age, glucose level, bmi,smoking status might be important variables for predicting

Multicollinearity and VIF

- Checked for multicollinearity of bmi on age, worktype, residence type, heart disease, hypertension, and glucose level. No severe multicollinearity was found



```
> vif(md)
                      GVIF Df GVIF^(1/(2*Df))
age               2.166315  1        1.471841
work_type         1.906632  4        1.084010
Residence_type    1.001376  1        1.000688
heart_disease     1.098335  1        1.048015
hypertension      1.102373  1        1.049939
avg_glucose_level 1.089470  1        1.043777
>
```

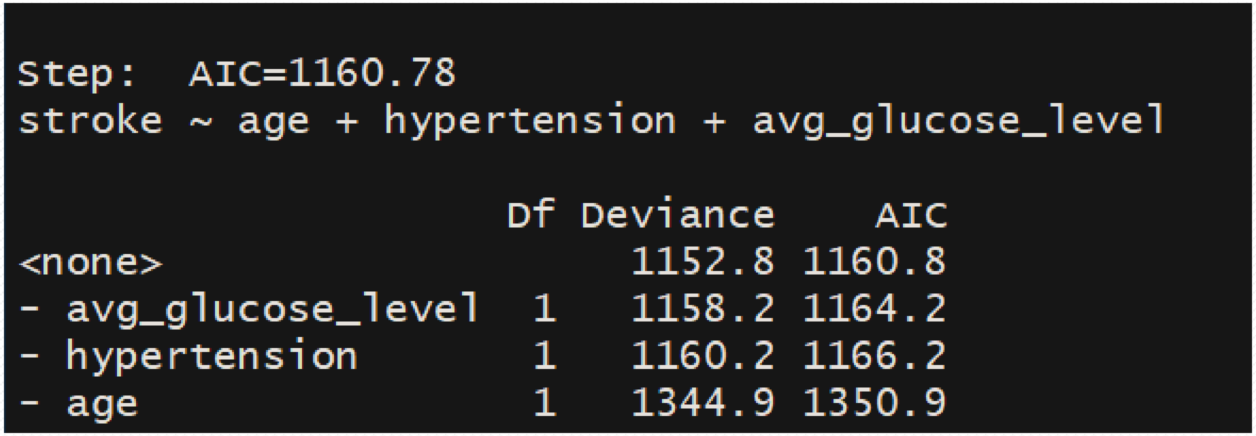
- Similarly, checking for multicollinearity between glucose and other variables, no severe multicollinearity was found
- Same can be said for hypertension, and age

Modelling - Preprocessing

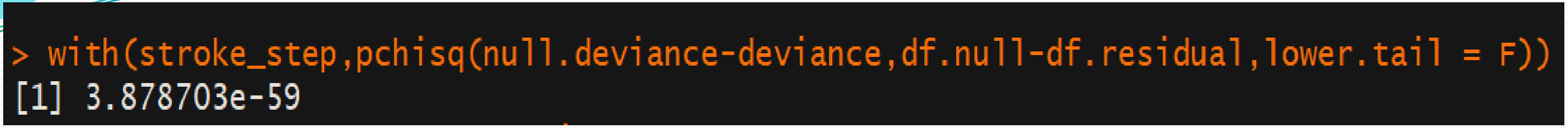
- Encoding the dummy variables of worktype, smoking status, and gender
- There are some outliers in bmi, and average glucose status. Not removing any datapoint as of now
- Doing a train-test split

Logistic Regression

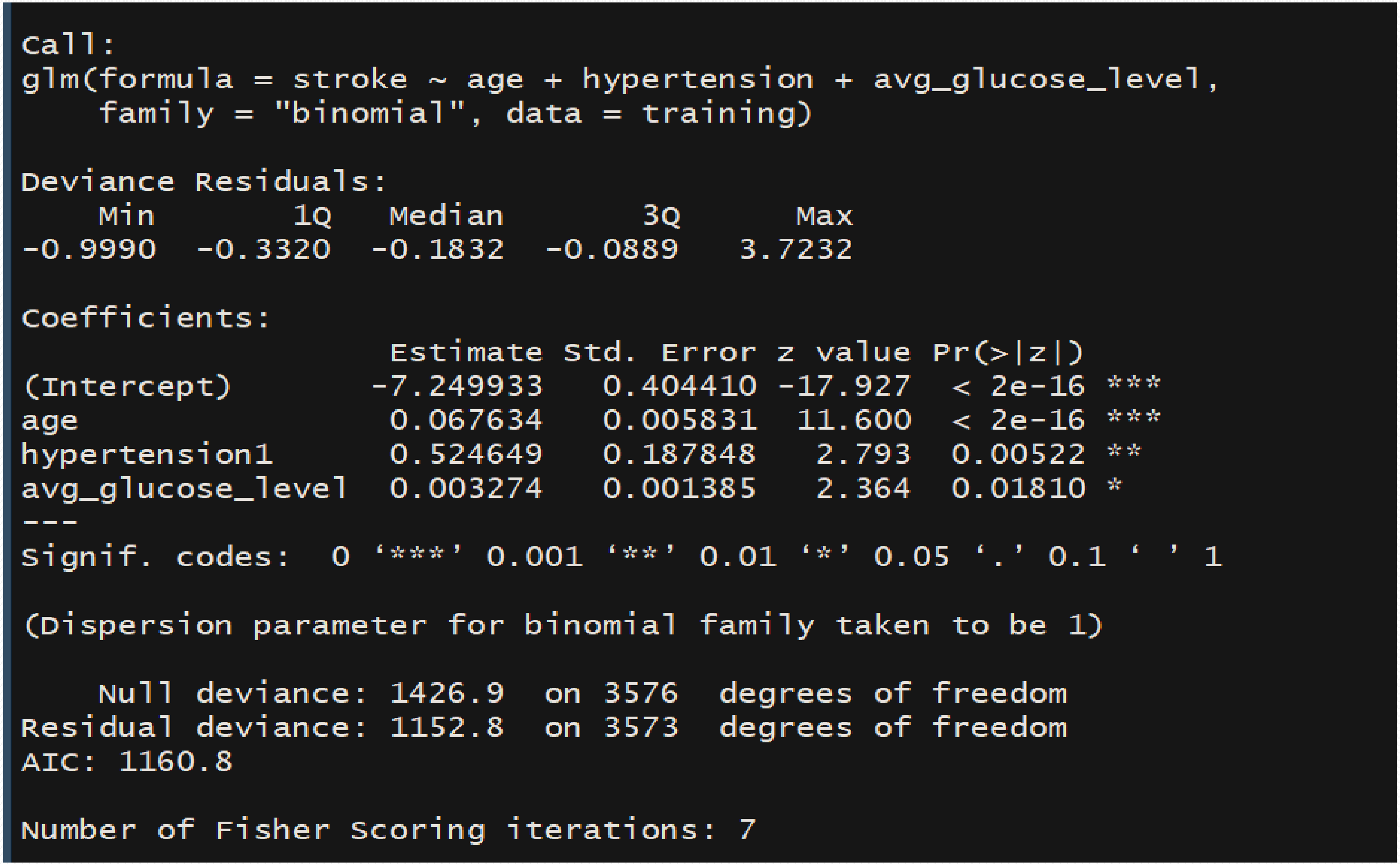
- After encoding the variables, had 15 variables. On performing a logistic regression, although the model was significant overall, most of the variables used were insignificant.
- So, I did a stepwise logistic regression.



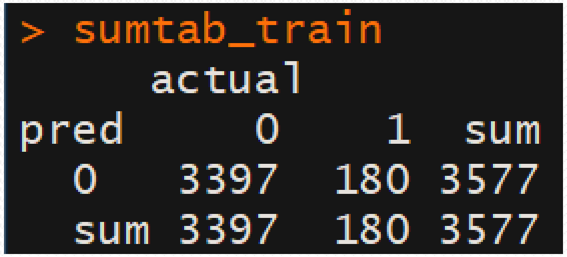
```
Step:  AIC=1160.78
stroke ~ age + hypertension + avg_glucose_level

                    Df Deviance    AIC
<none>                   1152.8 1160.8
- avg_glucose_level  1   1158.2 1164.2
- hypertension       1   1160.2 1166.2
- age                1   1344.9 1350.9
```



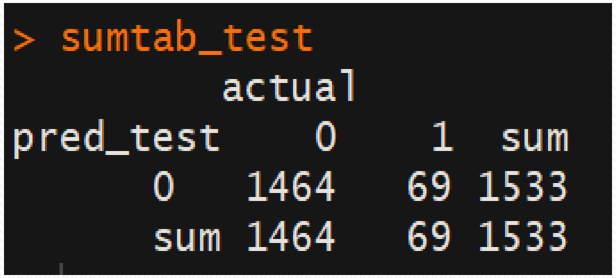
```
> with(stroke_step,pchisq(null.deviance-deviance,df.null-df.residual,lower.tail = F))
[1] 3.878703e-59
```



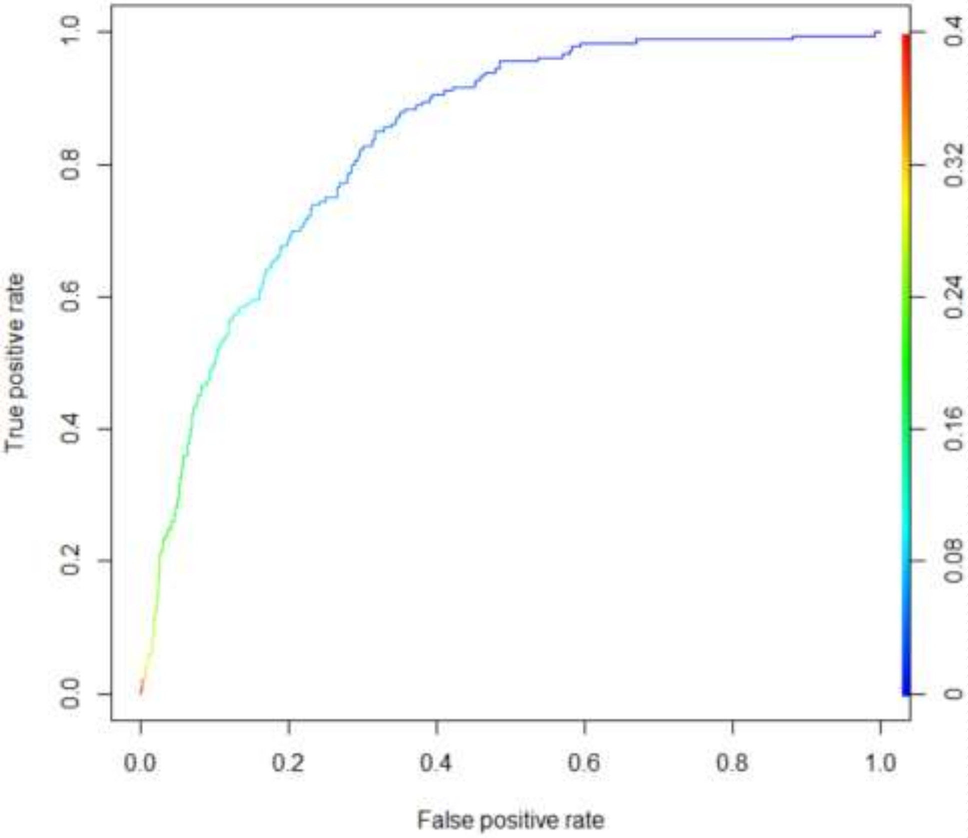
```
Call:
glm(formula = stroke ~ age + hypertension + avg_glucose_level, 
    family = "binomial", data = training)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.9990  -0.3320  -0.1832  -0.0889   3.7232  

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -7.249933   0.404410 -17.927  < 2e-16 ***
age                0.067634   0.005831  11.600  < 2e-16 ***
hypertension1      0.524649   0.187848   2.793  0.00522 ** 
avg_glucose_level  0.003274   0.001385   2.364  0.01810 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1426.9  on 3576  degrees of freedom
Residual deviance: 1152.8  on 3573  degrees of freedom
AIC: 1160.8

Number of Fisher Scoring iterations: 7
```

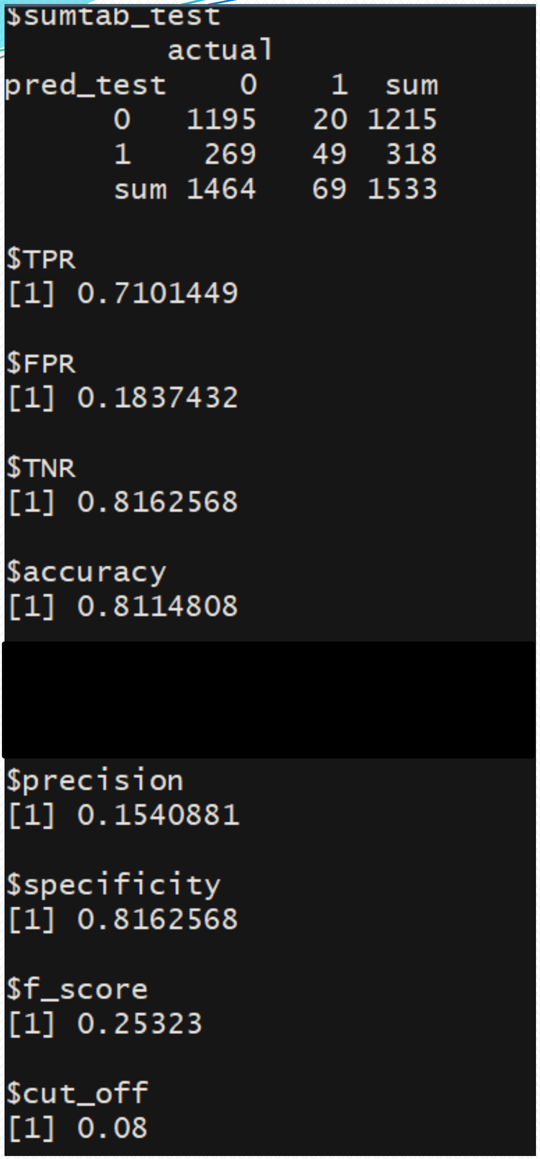


```
> sumtab_train
        actual
pred       0     1   sum
   0    3397   180  3577
   sum  3397   180  3577
```



```
> sumtab_test
          actual
pred_test     0     1   sum
        0  1464    69  1533
        sum 1464    69  1533
```

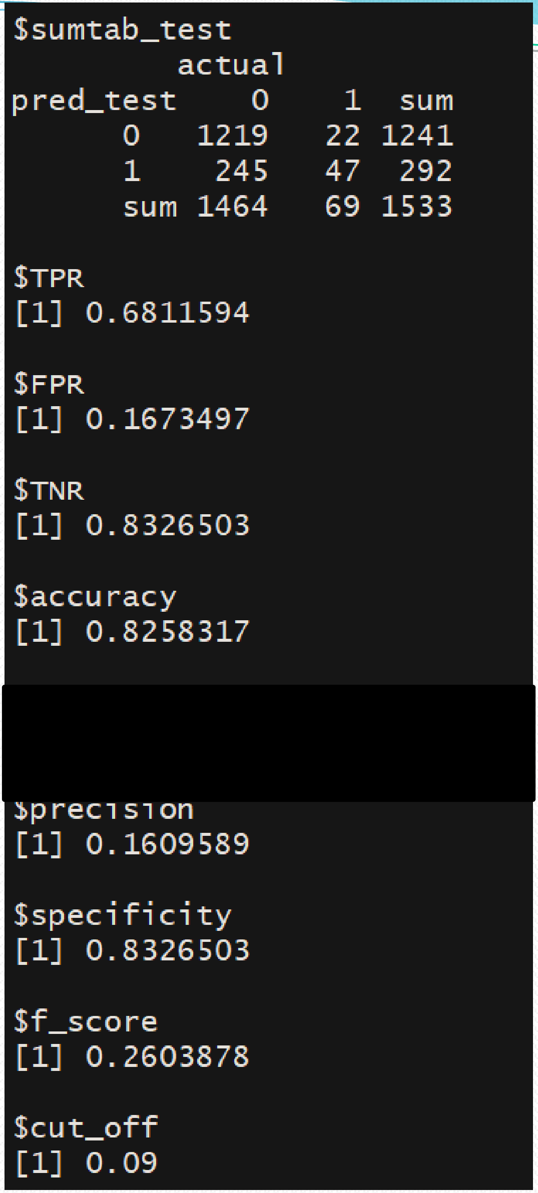




```
$sumtab_test
          actual
pred_test    0    1  sum
      0   1195   20 1215
      1    269   49  318
      sum 1464   69 1533

$TPR
[1] 0.7101449

$FPR
[1] 0.1837432

$TNR
[1] 0.8162568

$accuracy
[1] 0.8114808
```

[redacted]

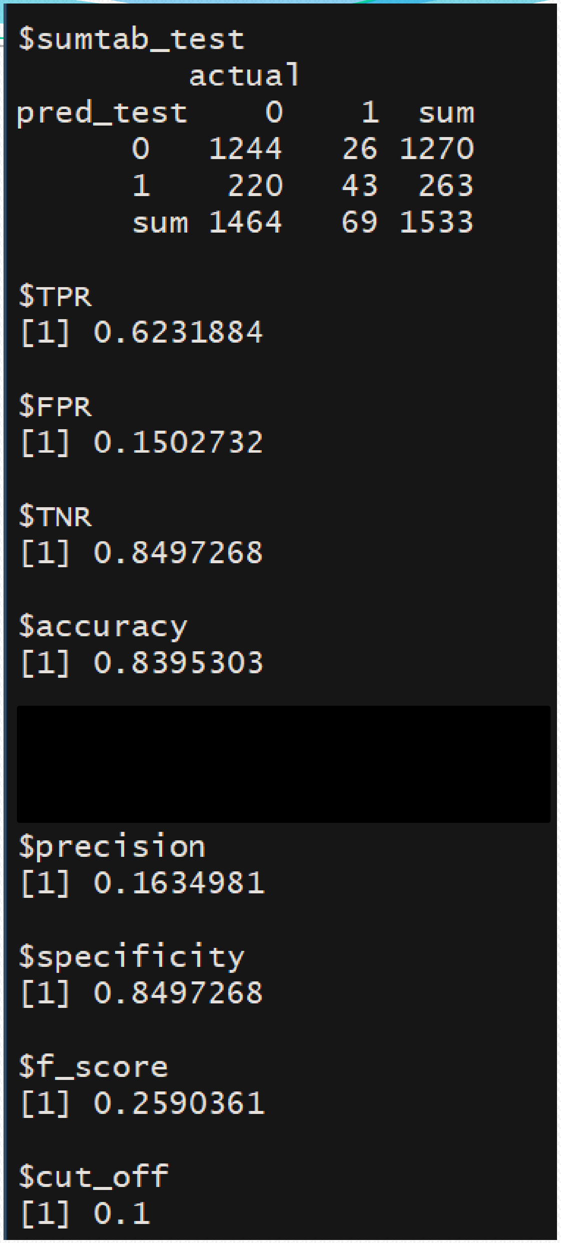
```
$precision
[1] 0.1540881

$specificity
[1] 0.8162568

$f_score
[1] 0.25323

$cut_off
[1] 0.08
```



```
$sumtab_test
          actual
pred_test    0    1  sum
      0   1219   22 1241
      1    245   47  292
      sum 1464   69 1533

$TPR
[1] 0.6811594

$FPR
[1] 0.1673497

$TNR
[1] 0.8326503

$accuracy
[1] 0.8258317
```

[redacted]

```
$precision
[1] 0.1609589

$specificity
[1] 0.8326503

$f_score
[1] 0.2603878

$cut_off
[1] 0.09
```



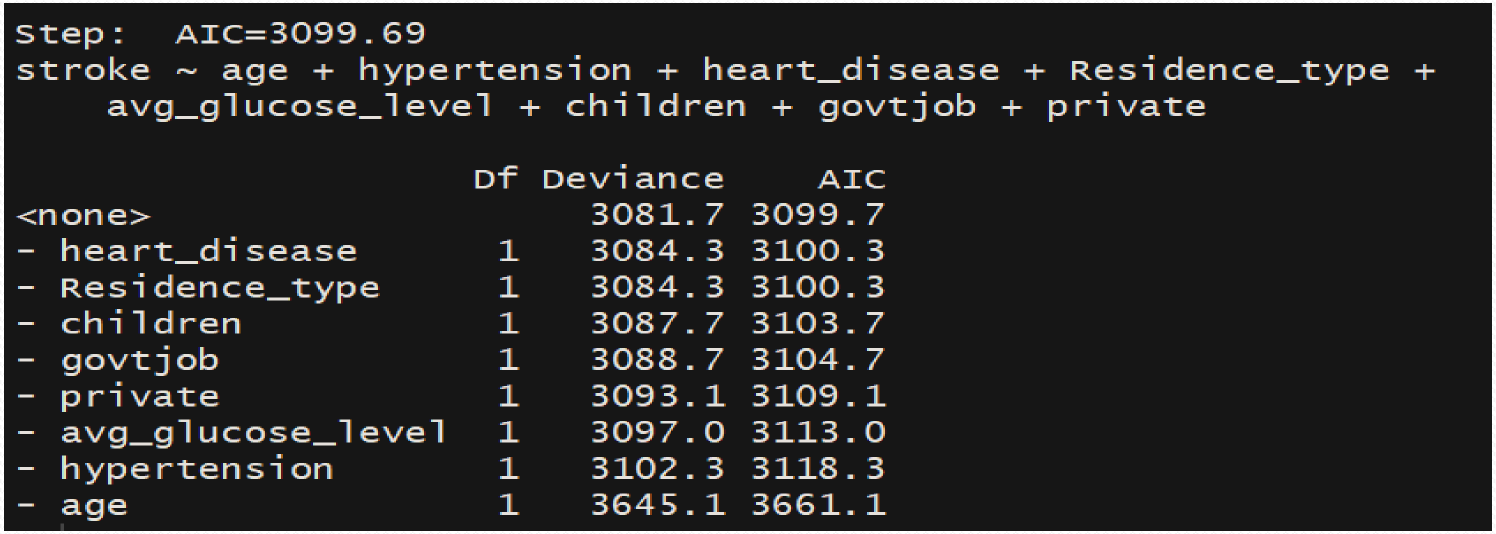
```
$sumtab_test
          actual
pred_test    0    1  sum
      0   1244   26 1270
      1    220   43  263
      sum 1464   69 1533

$TPR
[1] 0.6231884

$FPR
[1] 0.1502732

$TNR
[1] 0.8497268

$accuracy
[1] 0.8395303
```

[redacted]

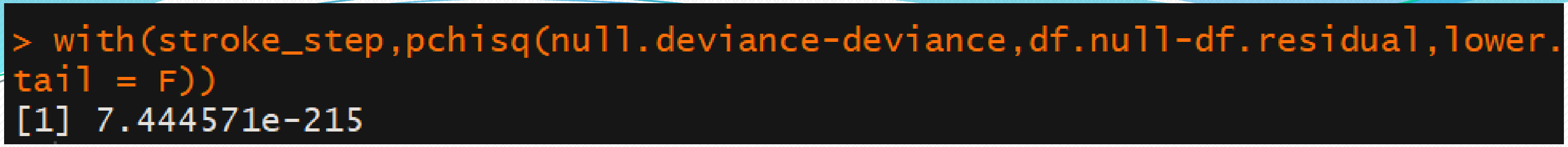
```
$precision
[1] 0.1634981

$specificity
[1] 0.8497268

$f_score
[1] 0.2590361

$cut_off
[1] 0.1
```

Oversampling

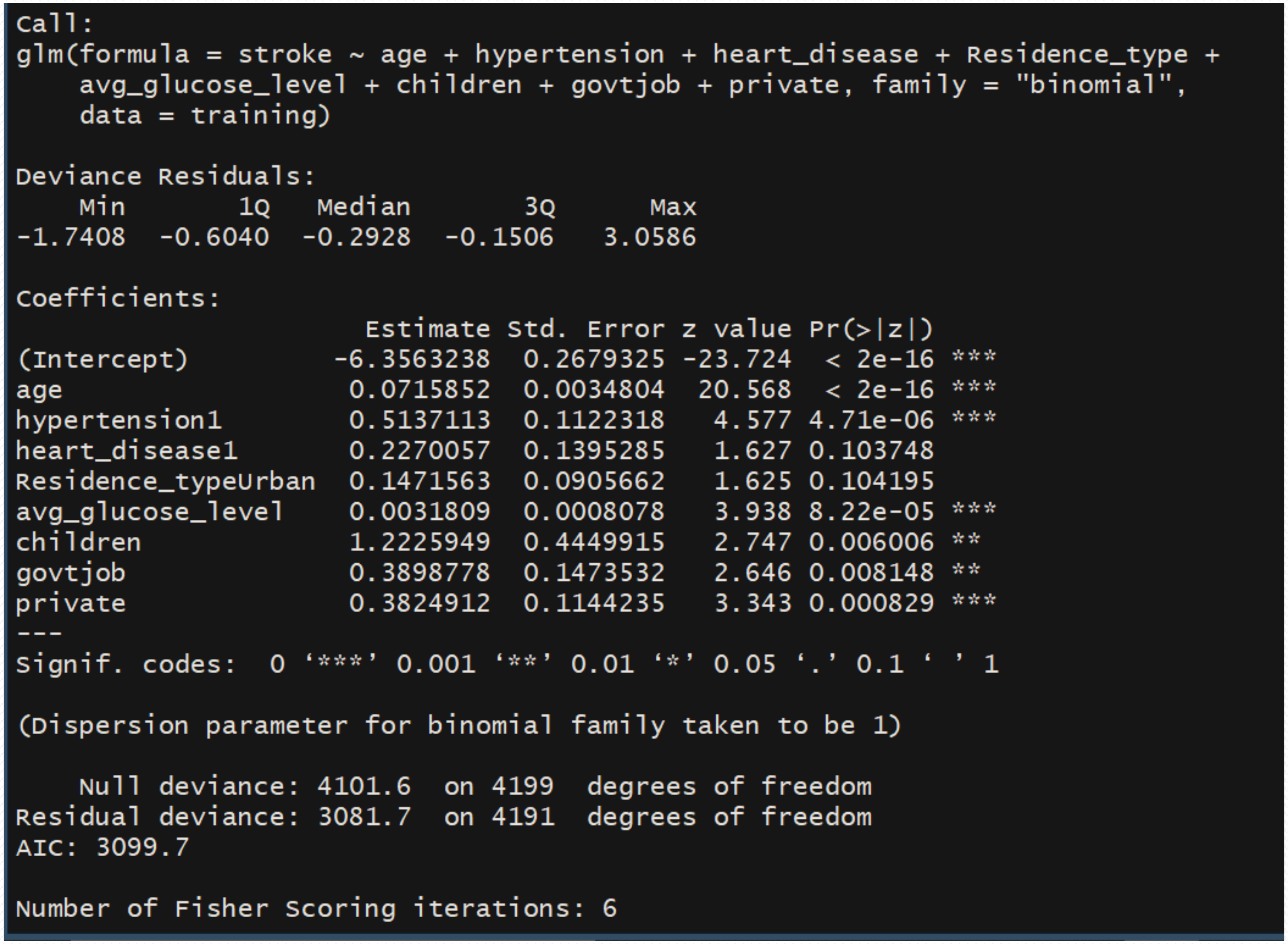
- Used ROSE package to oversample the data because of class imbalance
- After oversampling, the proportion of people who had stoke increased to around 19%
- Did stepwise regression on this oversampled data.



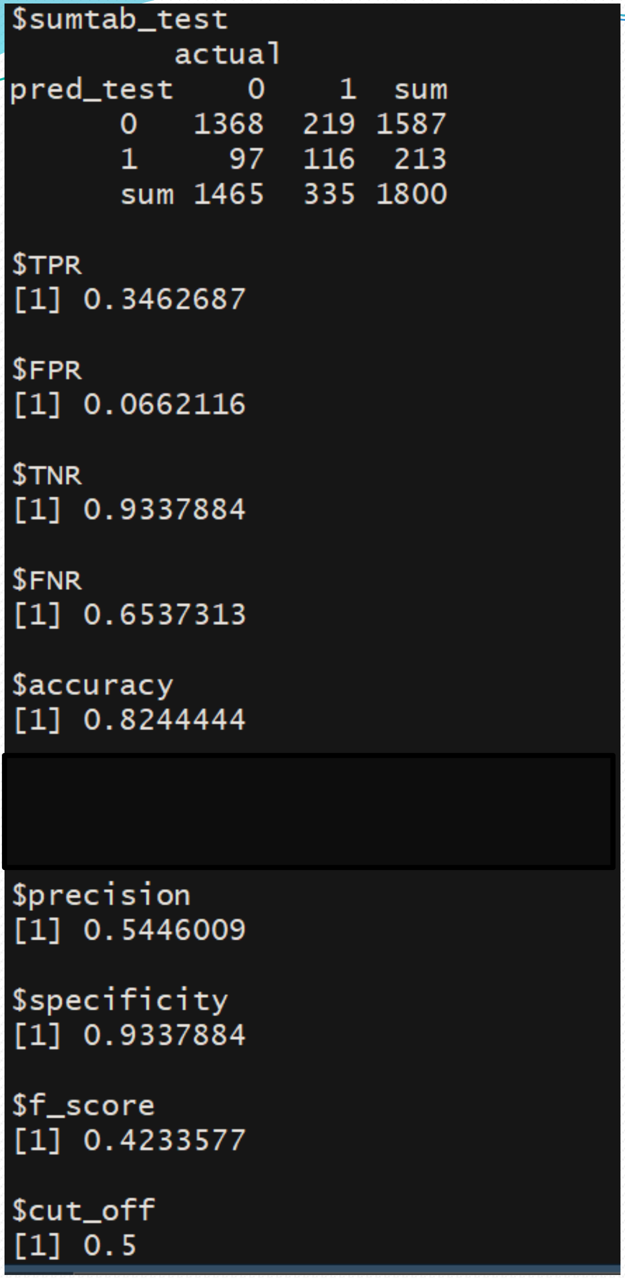
```
Step:  AIC=3099.69
stroke ~ age + hypertension + heart_disease + Residence_type +
    avg_glucose_level + children + govtjob + private

                    Df Deviance    AIC
<none>                   3081.7 3099.7
- heart_disease      1   3084.3 3100.3
- Residence_type     1   3084.3 3100.3
- children           1   3087.7 3103.7
- govtjob            1   3088.7 3104.7
- private            1   3093.1 3109.1
- avg_glucose_level  1   3097.0 3113.0
- hypertension       1   3102.3 3118.3
- age                1   3645.1 3661.1
```



```
> with(stroke_step,pchisq(null.deviance-deviance,df.null-df.residual,lower.
tail = F))
[1] 7.444571e-215
```



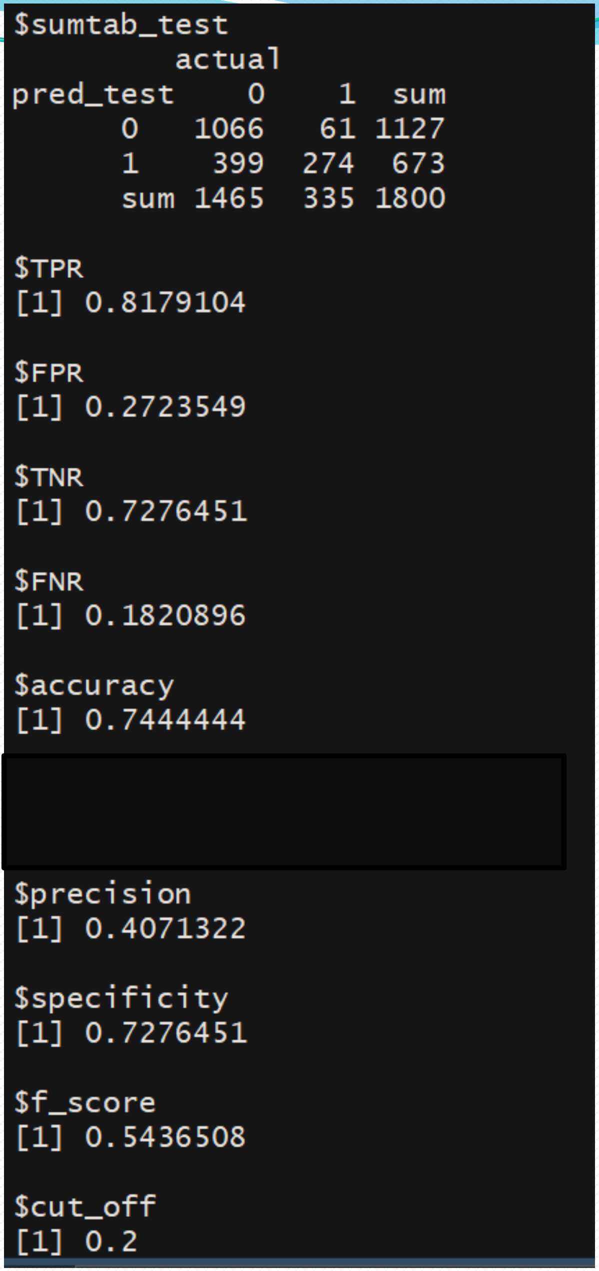
```
Call:
glm(formula = stroke ~ age + hypertension + heart_disease + Residence_type + 
    avg_glucose_level + children + govtjob + private, family = "binomial", 
    data = training)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7408  -0.6040  -0.2928  -0.1506   3.0586  

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -6.3563238  0.2679325 -23.724  < 2e-16 ***
age                  0.0715852  0.0034804  20.568  < 2e-16 ***
hypertension1        0.5137113  0.1122318   4.577 4.71e-06 ***
heart_disease1       0.2270057  0.1395285   1.627 0.103748    
Residence_typeUrban  0.1471563  0.0905662   1.625 0.104195    
avg_glucose_level    0.0031809  0.0008078   3.938 8.22e-05 ***
children             1.2225949  0.4449915   2.747 0.006006 ** 
govtjob              0.3898778  0.1473532   2.646 0.008148 ** 
private              0.3824912  0.1144235   3.343 0.000829 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4101.6  on 4199  degrees of freedom
Residual deviance: 3081.7  on 4191  degrees of freedom
AIC: 3099.7

Number of Fisher Scoring iterations: 6
```



```
$sumtab_test
         actual
pred_test    0    1  sum
      0   1368  219 1587
      1     97  116  213
      sum 1465  335 1800

$TPR
[1] 0.3462687

$FPR
[1] 0.0662116

$TNR
[1] 0.9337884

$FNR
[1] 0.6537313

$accuracy
[1] 0.8244444
```

[redacted]

```
$precision
[1] 0.5446009

$specificity
[1] 0.9337884

$f_score
[1] 0.4233577

$cut_off
[1] 0.5
```

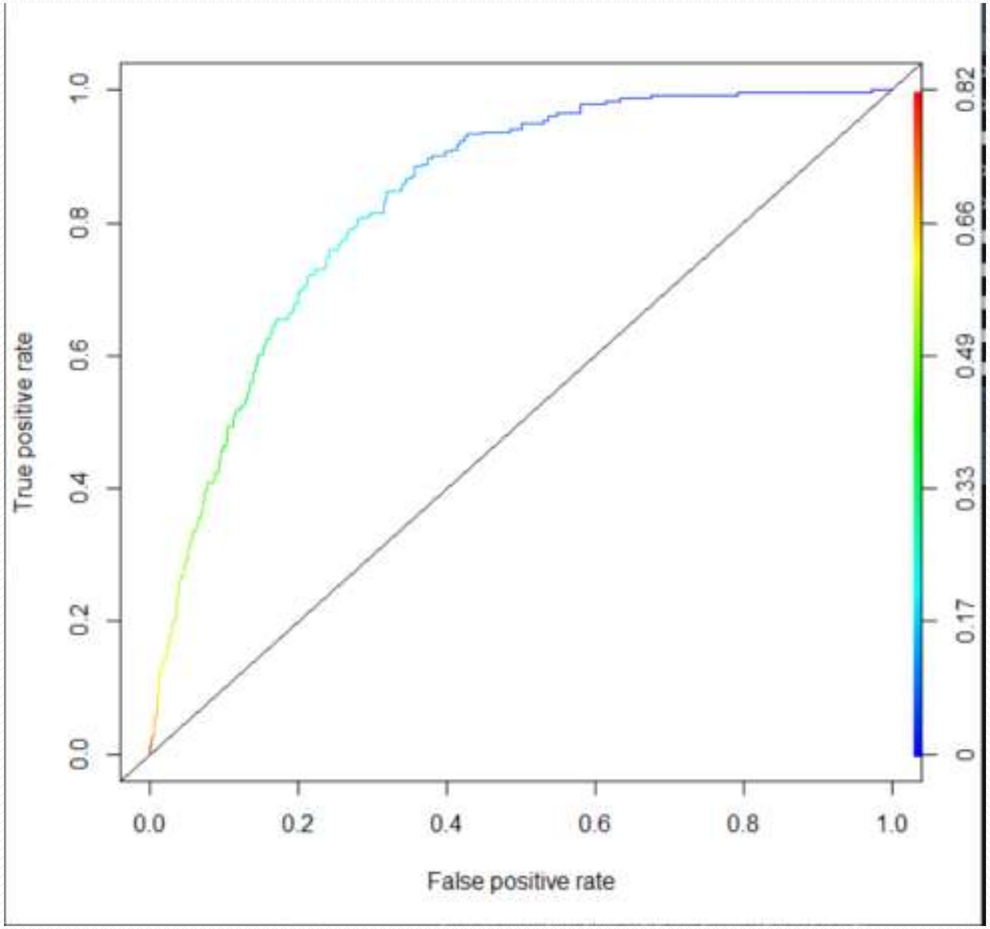


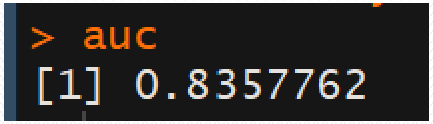
```
$sumtab_test
         actual
pred_test    0    1  sum
      0   1066   61 1127
      1    399  274  673
      sum 1465  335 1800

$TPR
[1] 0.8179104

$FPR
[1] 0.2723549

$TNR
[1] 0.7276451

$FNR
[1] 0.1820896

$accuracy
[1] 0.7444444
```

[redacted]

```
$precision
[1] 0.4071322

$specificity
[1] 0.7276451

$f_score
[1] 0.5436508

$cut_off
[1] 0.2
```

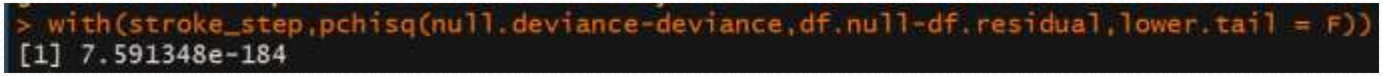
Roc curve, and AUC



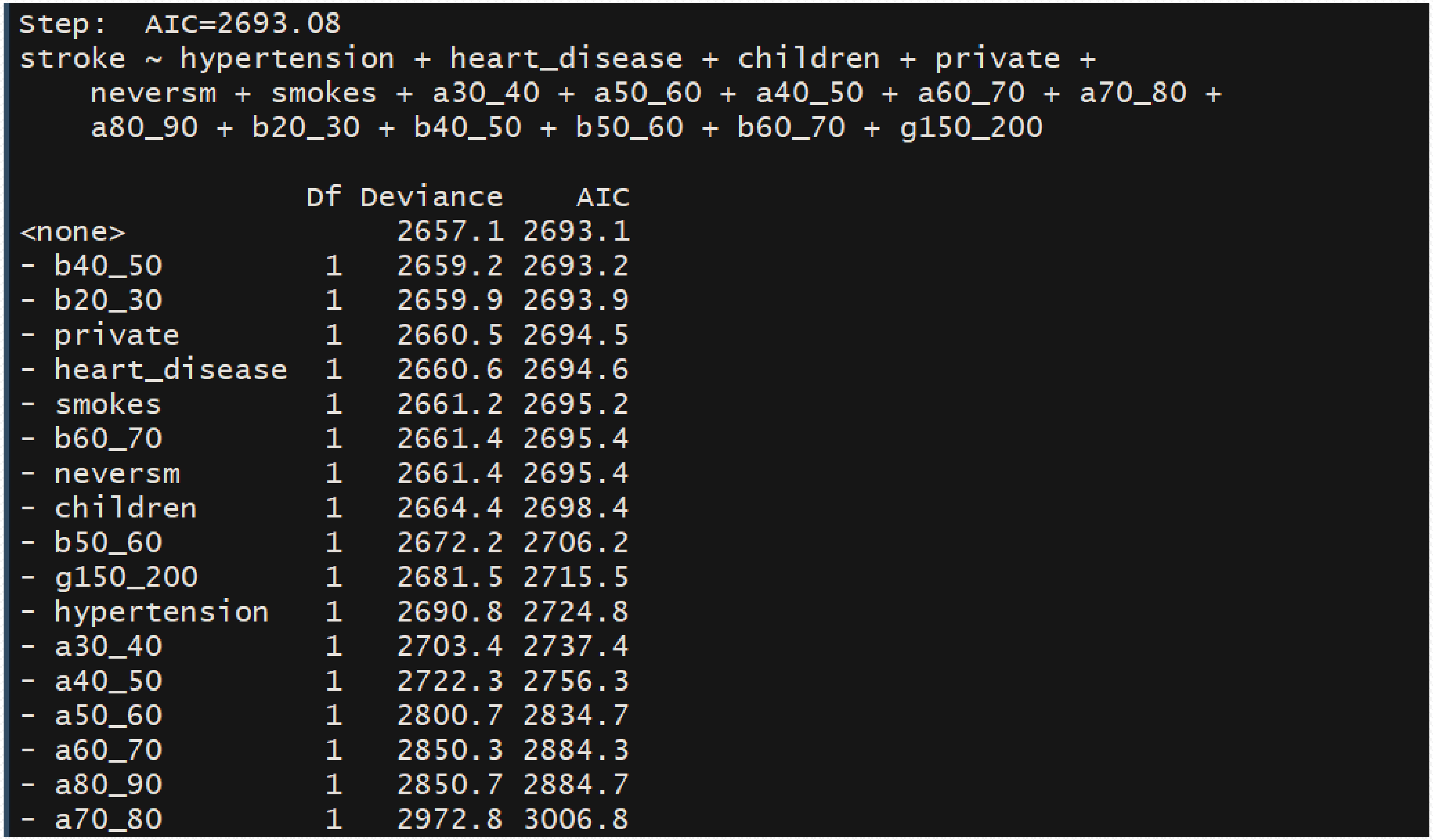


```
> auc
[1] 0.8357762
```

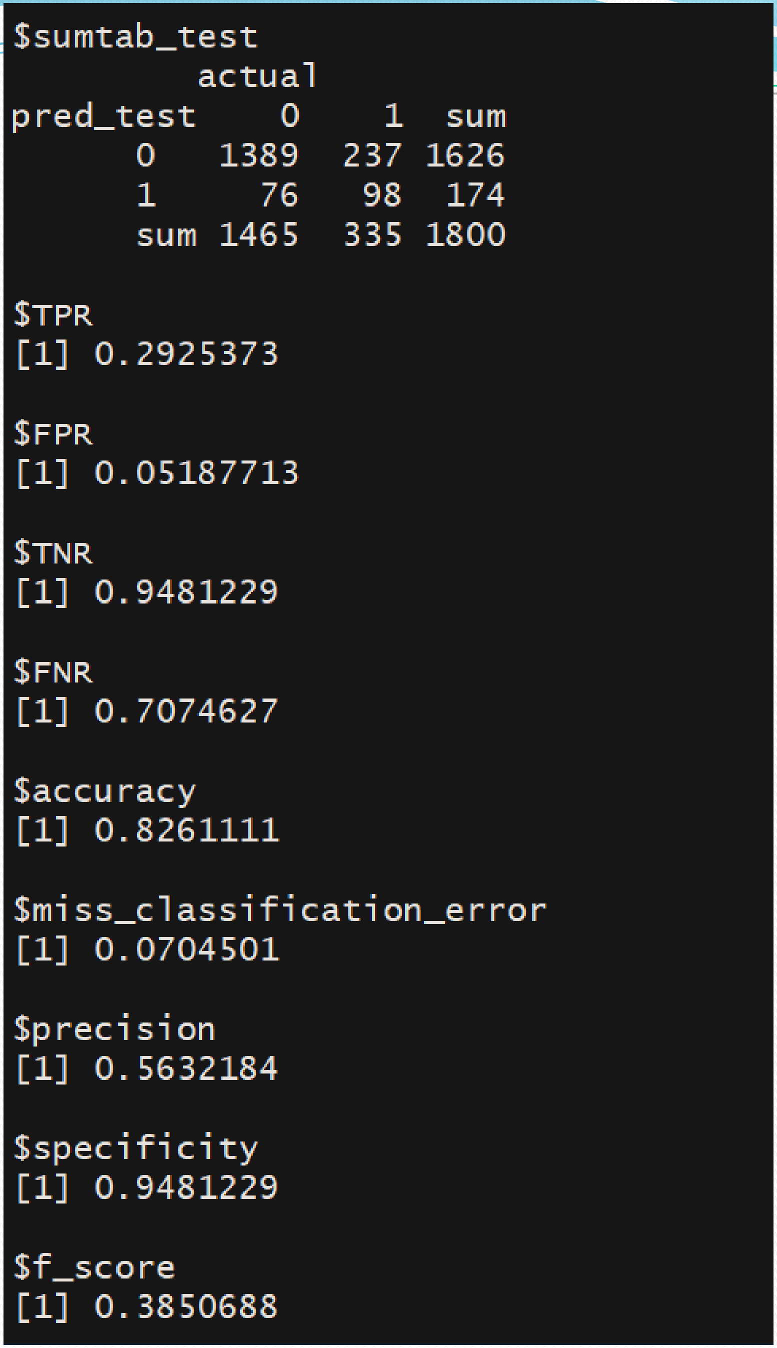
Logisitic Regression with binned numerical variables



```
> with(stroke_step,pchisq(null.deviance-deviance,df.null-df.residual,lower.tail = F))
[1] 7.591348e-184
```



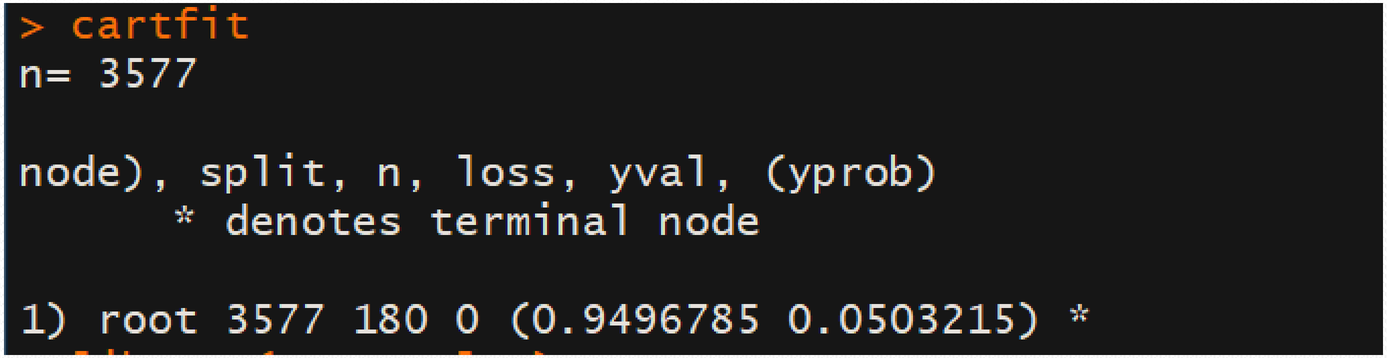
```
Step:  AIC=2693.08
stroke ~ hypertension + heart_disease + children + private + 
    neversm + smokes + a30_40 + a50_60 + a40_50 + a60_70 + a70_80 + 
    a80_90 + b20_30 + b40_50 + b50_60 + b60_70 + g150_200

                Df Deviance    AIC
<none>               2657.1 2693.1
- b40_50         1   2659.2 2693.2
- b20_30         1   2659.9 2693.9
- private        1   2660.5 2694.5
- heart_disease  1   2660.6 2694.6
- smokes         1   2661.2 2695.2
- b60_70         1   2661.4 2695.4
- neversm        1   2661.4 2695.4
- children       1   2664.4 2698.4
- b50_60         1   2672.2 2706.2
- g150_200       1   2681.5 2715.5
- hypertension   1   2690.8 2724.8
- a30_40         1   2703.4 2737.4
- a40_50         1   2722.3 2756.3
- a50_60         1   2800.7 2834.7
- a60_70         1   2850.3 2884.3
- a80_90         1   2850.7 2884.7
- a70_80         1   2972.8 3006.8
```



```
$sumtab_test
         actual
pred_test    0    1  sum
      0   1389  237 1626
      1     76   98  174
      sum 1465  335 1800

$TPR
[1] 0.2925373

$FPR
[1] 0.05187713

$TNR
[1] 0.9481229

$FNR
[1] 0.7074627

$accuracy
[1] 0.8261111

$miss_classification_error
[1] 0.0704501

$precision
[1] 0.5632184

$specificity
[1] 0.9481229

$f_score
[1] 0.3850688
```

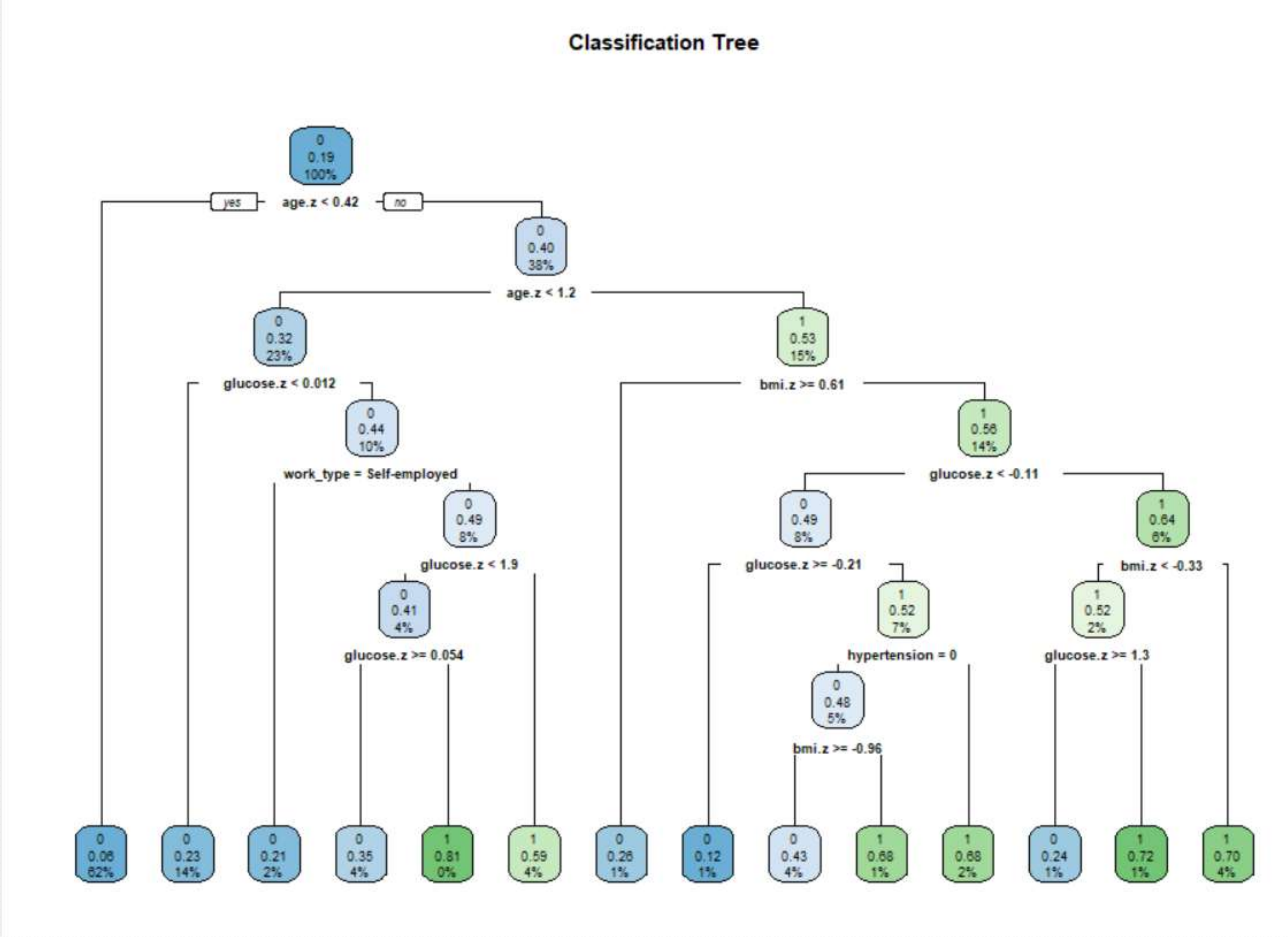
Tree based models

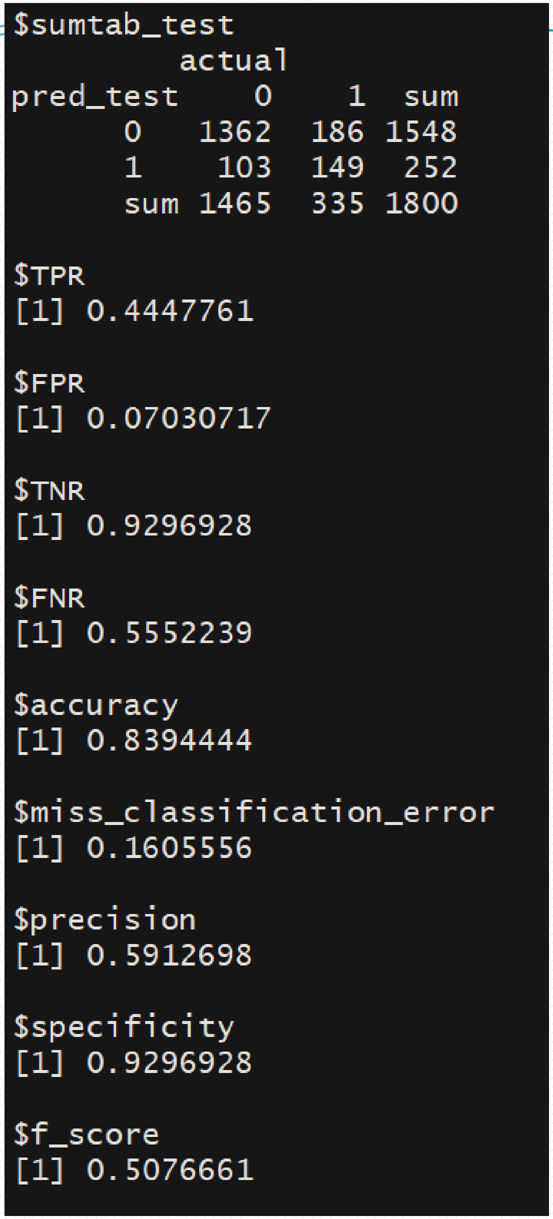
- Normalized the numerical variables
- CART model on original data (data not oversampled)



```
> cartfit
n= 3577

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 3577 180 0 (0.9496785 0.0503215) *
```

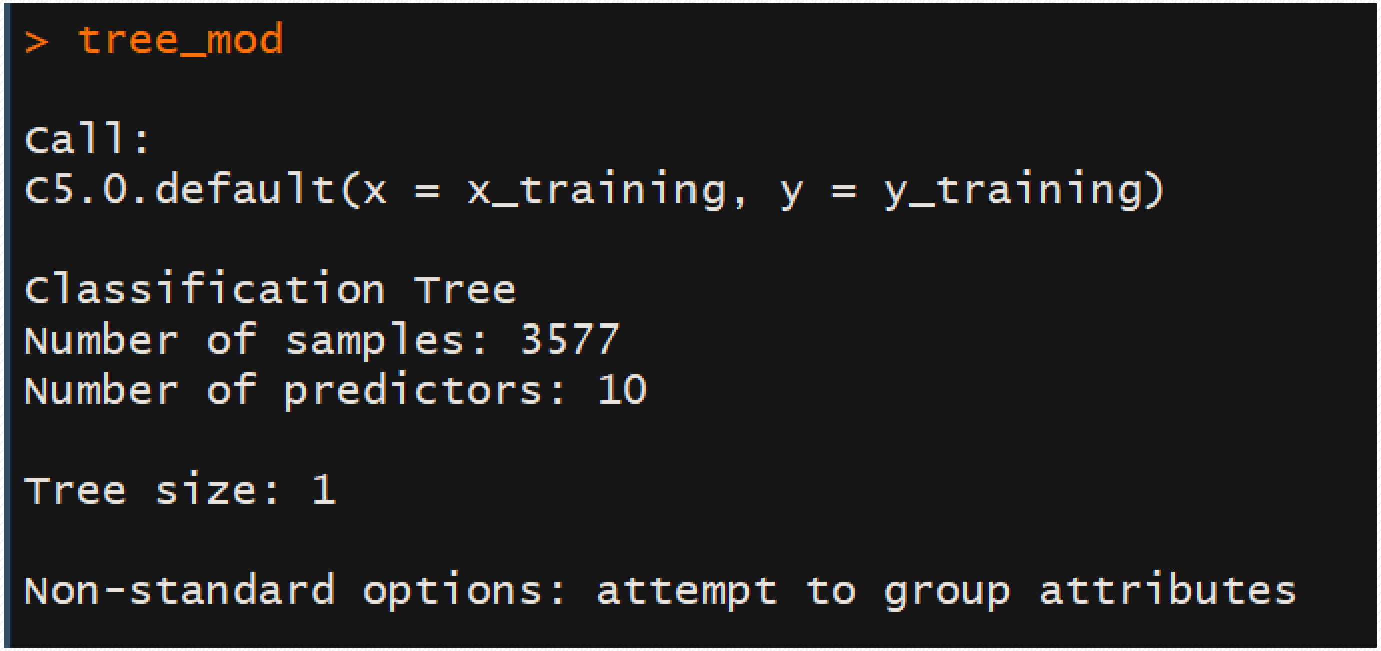
CART on the oversampled data





```
$sumtab_test
         actual
pred_test    0    1  sum
      0   1362  186 1548
      1    103  149  252
      sum 1465  335 1800

$TPR
[1] 0.4447761

$FPR
[1] 0.07030717

$TNR
[1] 0.9296928

$FNR
[1] 0.5552239

$accuracy
[1] 0.8394444

$miss_classification_error
[1] 0.1605556

$precision
[1] 0.5912698

$specificity
[1] 0.9296928

$f_score
[1] 0.5076661
```

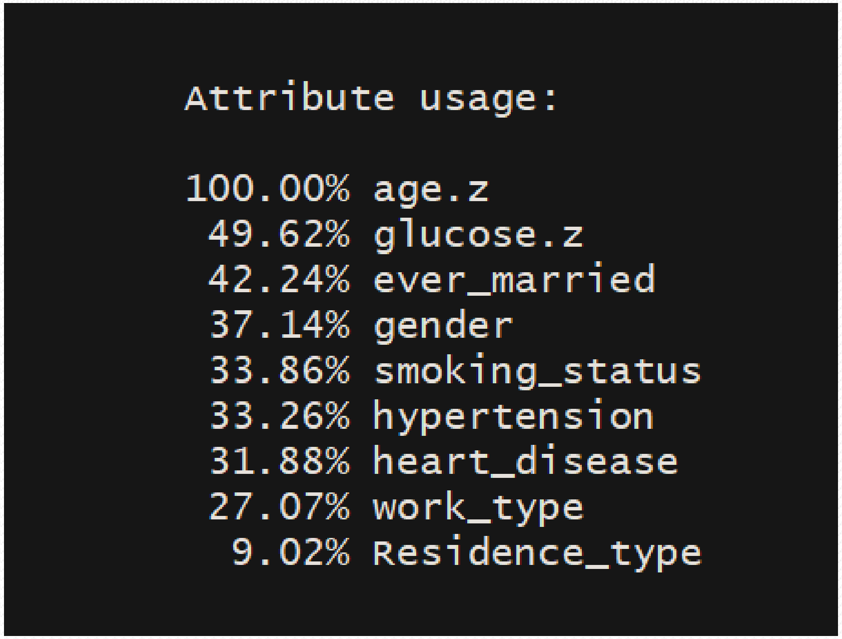
C5 tree on data not oversampled



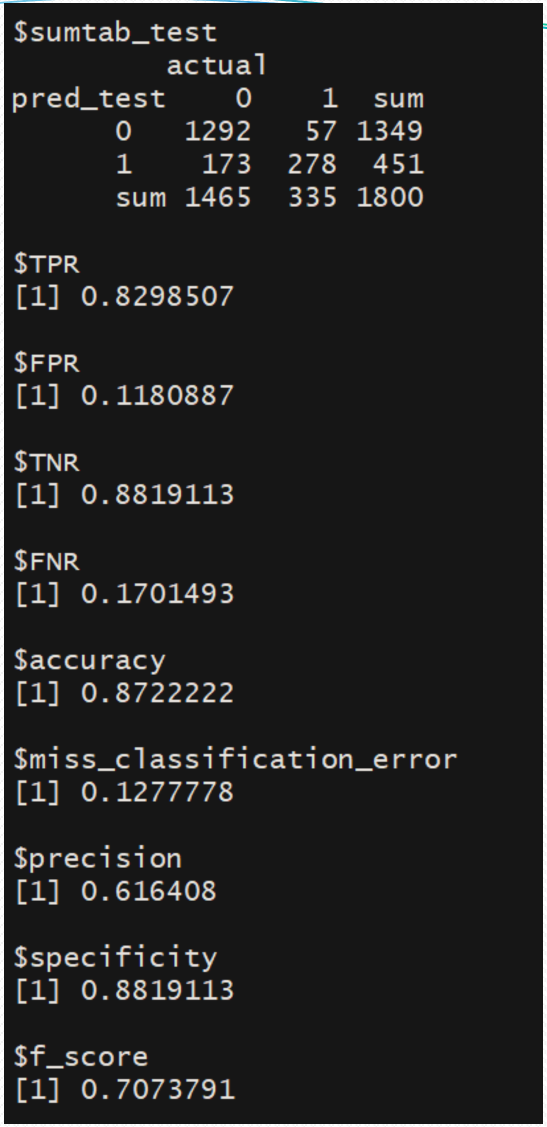
```
> tree_mod

Call:
C5.0.default(x = x_training, y = y_training)

Classification Tree
Number of samples: 3577
Number of predictors: 10

Tree size: 1

Non-standard options: attempt to group attributes
```

C5 tree on oversampled data

- The tree was big. The attributes used were

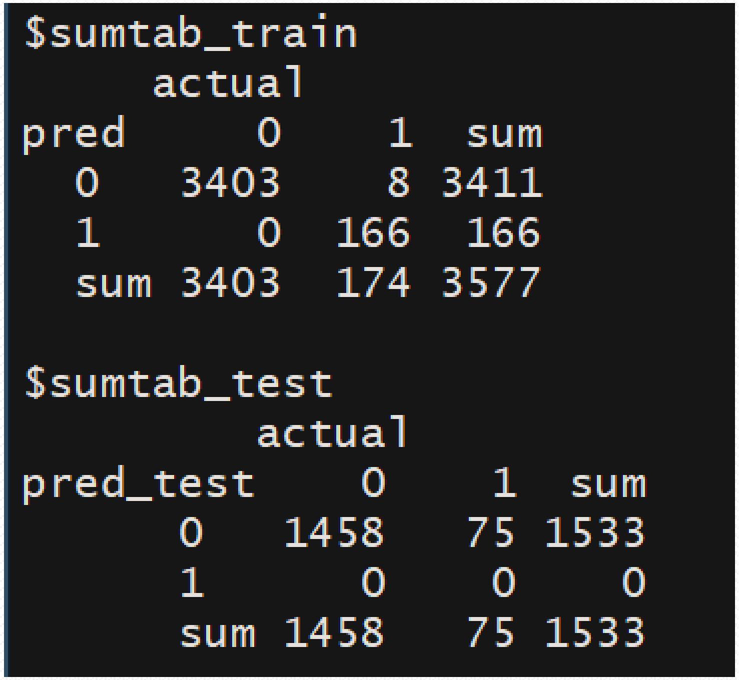


```
Attribute usage:

100.00% age.z
 49.62% glucose.z
 42.24% ever_married
 37.14% gender
 33.86% smoking_status
 33.26% hypertension
 31.88% heart_disease
 27.07% work_type
  9.02% Residence_type
```



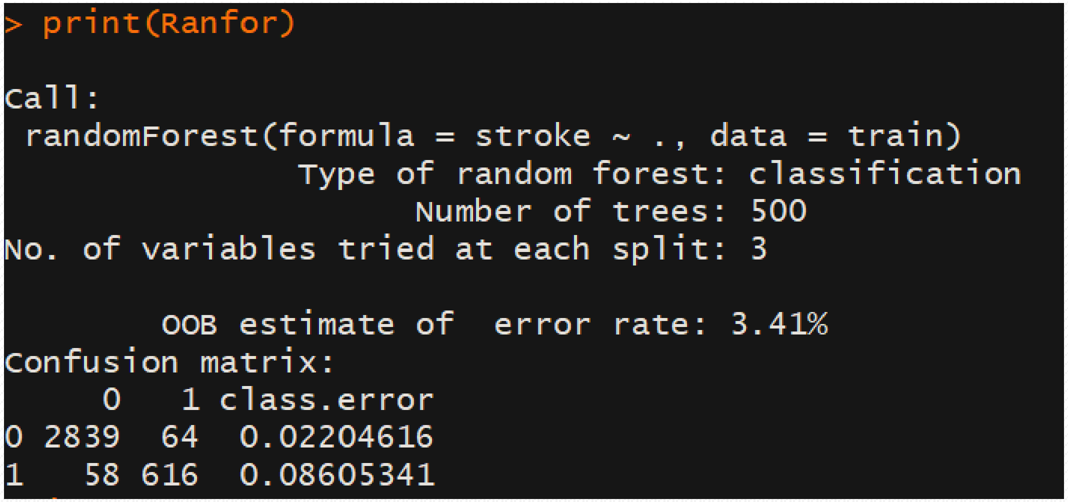
```
$sumtab_test
         actual
pred_test    0    1  sum
      0   1292   57 1349
      1    173  278  451
      sum 1465  335 1800

$TPR
[1] 0.8298507

$FPR
[1] 0.1180887

$TNR
[1] 0.8819113

$FNR
[1] 0.1701493

$accuracy
[1] 0.8722222

$miss_classification_error
[1] 0.1277778

$precision
[1] 0.616408

$specificity
[1] 0.8819113

$f_score
[1] 0.7073791
```

RF on data not oversampled

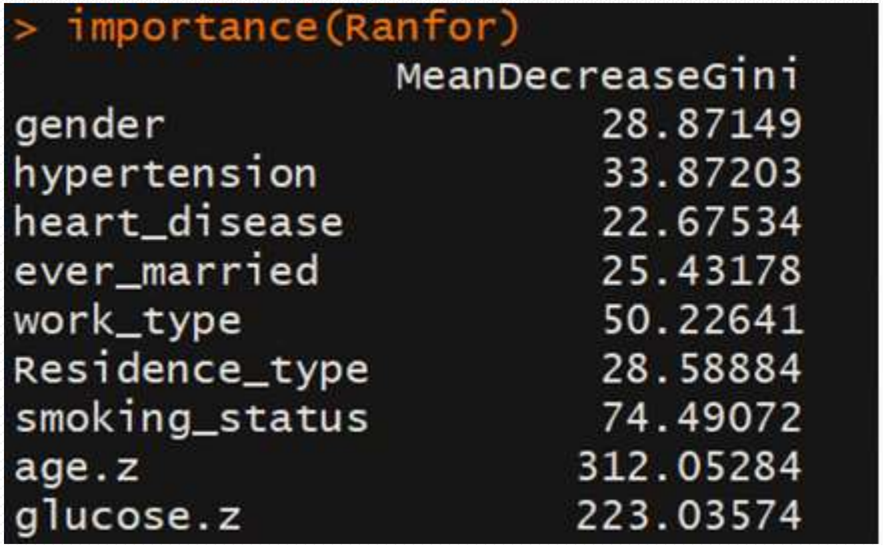


```
$sumtab_train
     actual
pred     0    1  sum
  0   3403    8 3411
  1      0  166  166
  sum 3403  174 3577

$sumtab_test
         actual
pred_test    0    1  sum
      0   1458   75 1533
      1      0    0    0
      sum 1458   75 1533
```

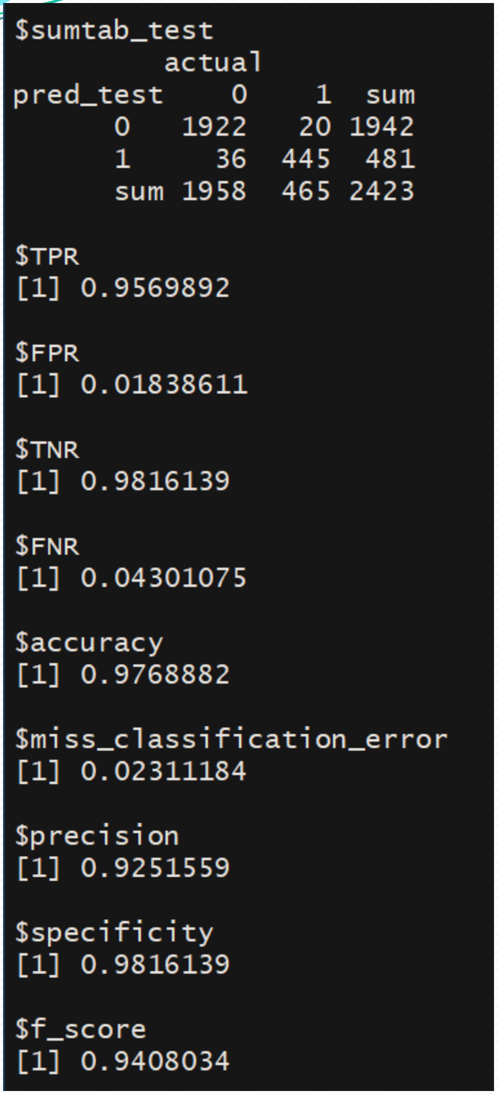
Random Forest on oversampled data



```
> print(Ranfor)

Call:
 randomForest(formula = stroke ~ ., data = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 3.41%
Confusion matrix:
     0   1 class.error
0 2839  64  0.02204616
1   58 616  0.08605341
```

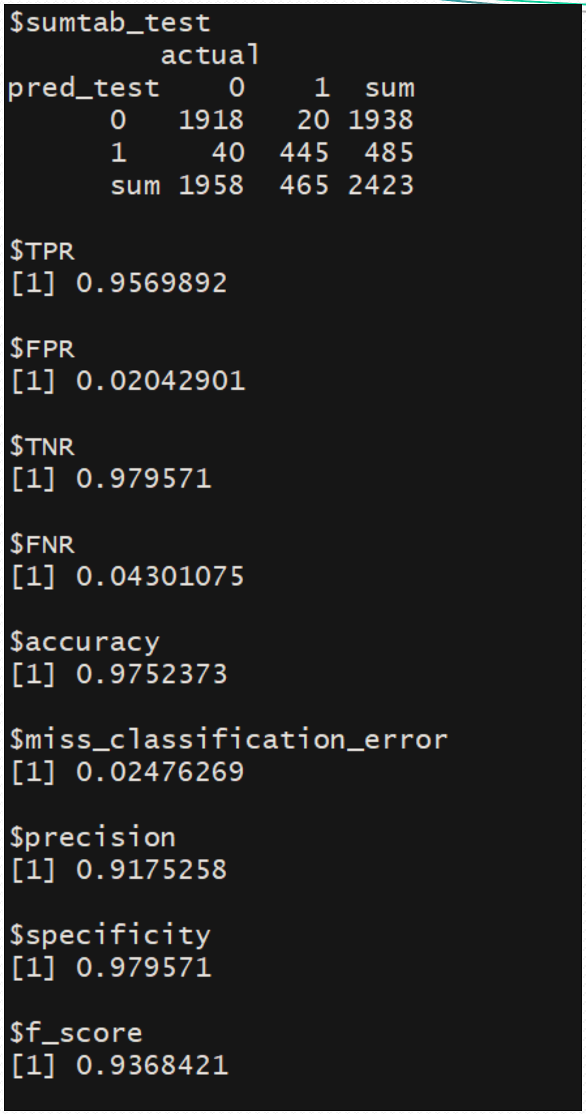


```
> importance(Ranfor)
                  MeanDecreaseGini
gender                    28.87149
hypertension              33.87203
heart_disease             22.67534
ever_married              25.43178
work_type                 50.22641
Residence_type            28.58884
smoking_status            74.49072
age.z                    312.05284
glucose.z                223.03574
```



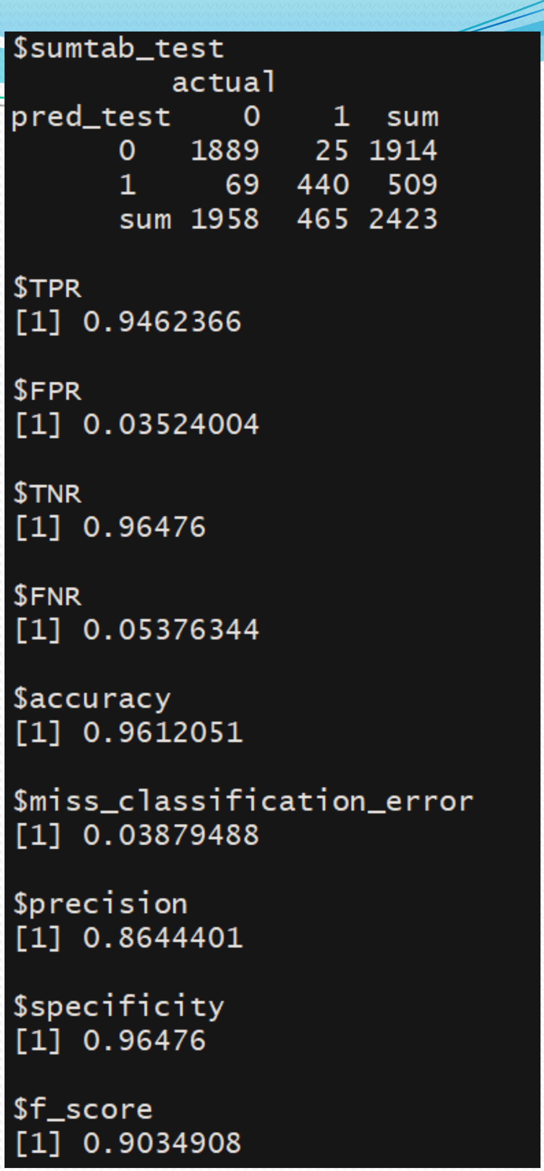
```
$sumtab_test
        actual
pred_test    0    1  sum
      0   1922   20 1942
      1     36  445  481
      sum 1958  465 2423

$TPR
[1] 0.9569892

$FPR
[1] 0.01838611

$TNR
[1] 0.9816139

$FNR
[1] 0.04301075

$accuracy
[1] 0.9768882

$miss_classification_error
[1] 0.02311184

$precision
[1] 0.9251559

$specificity
[1] 0.9816139

$f_score
[1] 0.9408034
```

All variables used



```
$sumtab_test
        actual
pred_test    0    1  sum
      0   1918   20 1938
      1     40  445  485
      sum 1958  465 2423

$TPR
[1] 0.9569892

$FPR
[1] 0.02042901

$TNR
[1] 0.979571

$FNR
[1] 0.04301075

$accuracy
[1] 0.9752373

$miss_classification_error
[1] 0.02476269

$precision
[1] 0.9175258

$specificity
[1] 0.979571

$f_score
[1] 0.9368421
```

Removing
Gender



```
$sumtab_test
        actual
pred_test    0    1  sum
      0   1889   25 1914
      1     69  440  509
      sum 1958  465 2423

$TPR
[1] 0.9462366

$FPR
[1] 0.03524004

$TNR
[1] 0.96476

$FNR
[1] 0.05376344

$accuracy
[1] 0.9612051

$miss_classification_error
[1] 0.03879488

$precision
[1] 0.8644401

$specificity
[1] 0.96476

$f_score
[1] 0.9034908
```

Glucose, age, smoking
status, and work type

Thank You