Statistical Analysis Report

Author: Shaarang Buckal

Background

The following analysis is performed to decrease the time lag in diagnosis of Autism. The urgency of Autism diagnosis is cardinal to reduce the fiscal and ethical costs associated with the medical condition.

Data Source

The Data is a subset of data from Dr. Fadi Fayez Thabtah, Department of Digital Technology, Manukau Institute of Technology, Auckland, New Zealand.

Data Transformation and Cleaning (Description)

Autism

Autism variable originally marks the presence or absence of autism in a single case, it is a character variable. It was converted into a binary variable named Autism_Bin (it is a column name hence the name initials were not appended), Yes being replaced by 1 and No being replaced by 0.

Nationality

The variable identifying the patient's nationality/continental origin was transformed to six dummy variables.

Rel

The variable identifying the professional or the person who completed the test was transformed to five dummy variables.

Descriptive Data Analysis

```
     Others            Parent           Relative            Self           African      
 Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.0000   Median :0.00000   Median :0.00000   Median :1.0000   Median :0.0000  
 Mean   :0.1303   Mean   :0.07415   Mean   :0.03808   Mean   :0.7415   Mean   :0.1443  
 3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000  
 Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000  
     Asian           European       LatinAmerica    MiddleEastern    NorthAmerican    
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000  
 Mean   :0.1583   Mean   :0.1483   Mean   :0.1483   Mean   :0.1603   Mean   :0.2405  
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
```

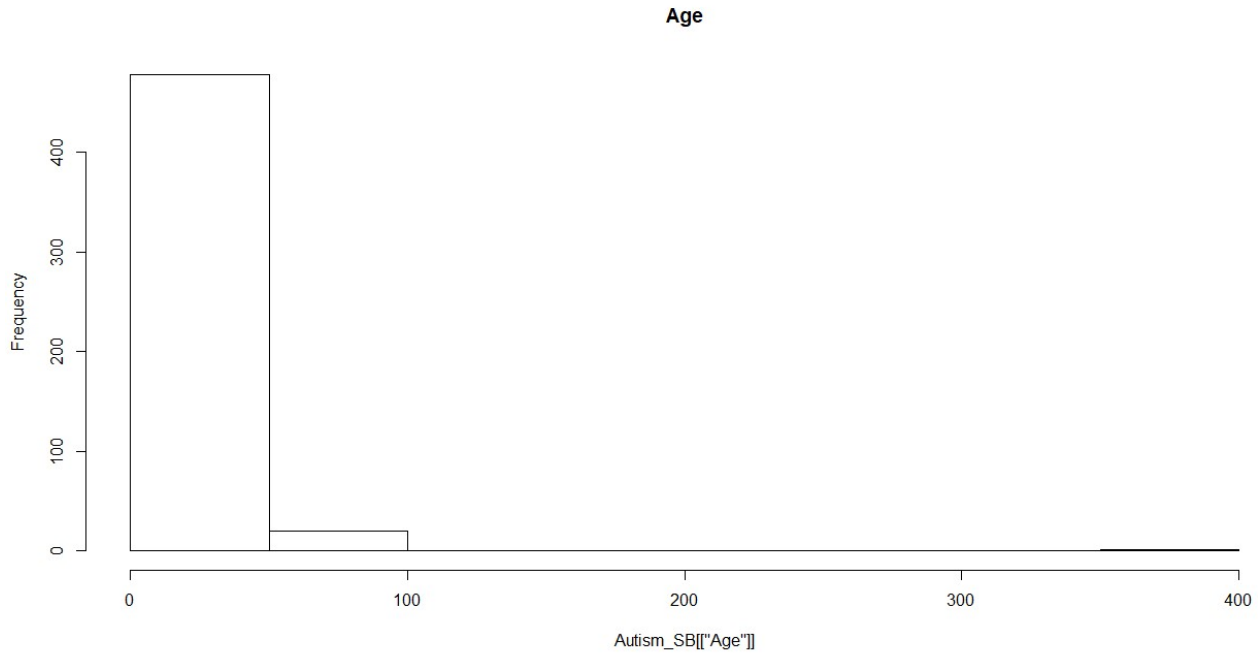
```
      A01              A02              A03              A04             A05       
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.000  
 Median :1.0000   Median :0.0000   Median :0.0000   Median :0.000   Median :0.000  
 Mean   :0.7255   Mean   :0.4449   Mean   :0.4589   Mean   :0.495   Mean   :0.497  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.000  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   :1.000  
      A06              A07              A08              A09             A10       
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.000  
 Median :0.0000   Median :0.0000   Median :1.0000   Median :1.000   Median :1.000  
 Mean   :0.2886   Mean   :0.4108   Mean   :0.6212   Mean   :0.507   Mean   :0.505  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.000  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   :1.000  
      Age              Gender           Jaundic          Autism_Bin      HealthCareProf    
 Min.   : 17.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000  
 1st Qu.: 21.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000  
 Median : 27.00   Median :1.0000   Median :0.0000   Median :0.0000   Median :0.00000  
 Mean   : 29.91   Mean   :0.5251   Mean   :0.1062   Mean   :0.2585   Mean   :0.01603  
 3rd Qu.: 35.00   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.00000  
 Max.   :383.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000  
```

All the variables are binary except for the Age variable. For binary variables, the descriptive summary does not add any insight, anyhow the descriptive summary for binary variables is as expected.

The only non-binary variable Age displays a clear discrepancy in the data, maximum age recorded for a case is 383 years, as much as we humans would want to outlive our biological clock this is clearly a mistake in data input. The second oldest case is 61 years old; I was in a dilemma whether to correct it as another 61 or 62, I chose the former.

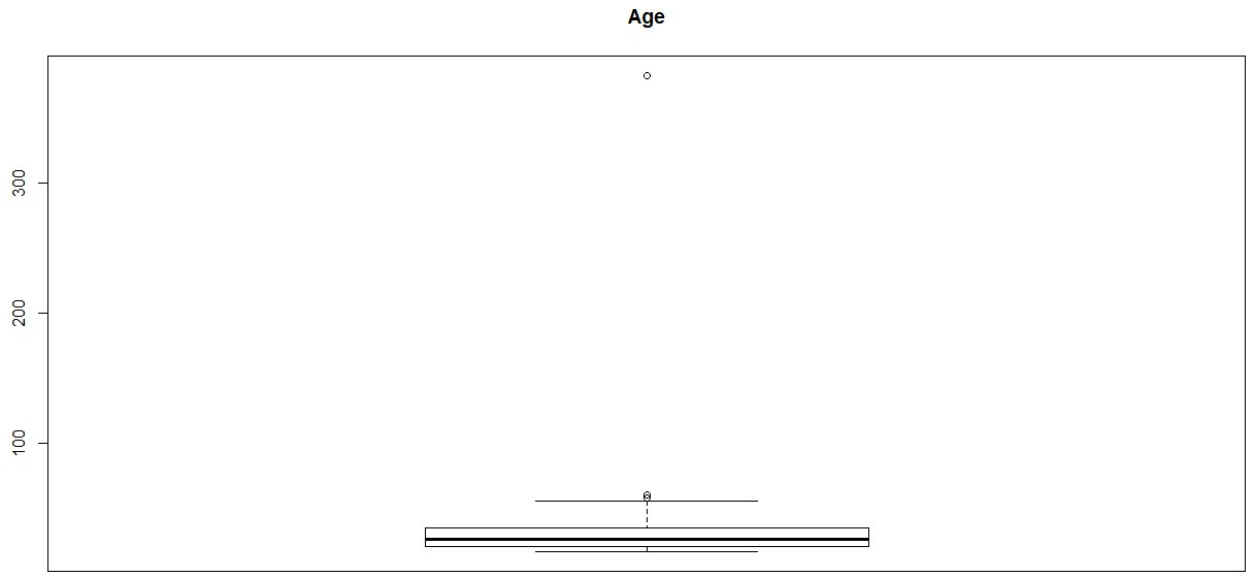
Like the descriptive summaries, histograms for binary variables is illogical in my humble opinion, whatever insight we could harness from binary variable, it would be done in the logistic regression model.

Below is the histogram for non-binary age variable.



The single outlier of 383 has distorted the histogram scale, the single protrusion at the far right-end is the outlier.

Outlier



Like the histogram section I have not produced the boxplot for binary variables, moving to the more pertinent Age variable, the boxplot displays clearly displays the effect of a single outlier. We shall see the boxplot and histograms below, after handling the outlier.

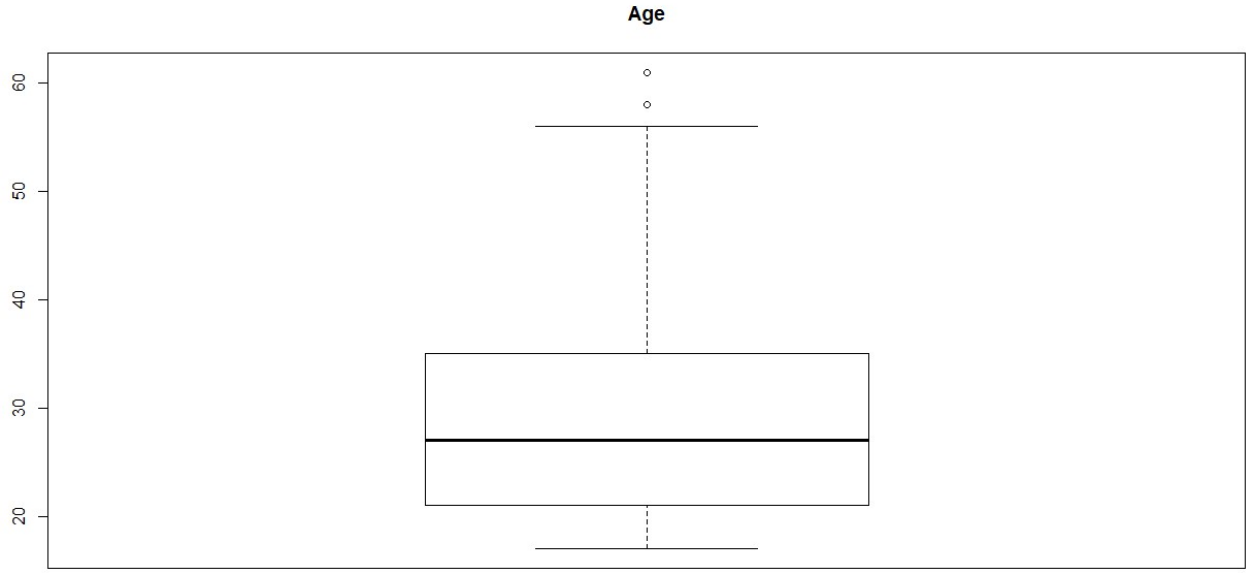
Outlier handling

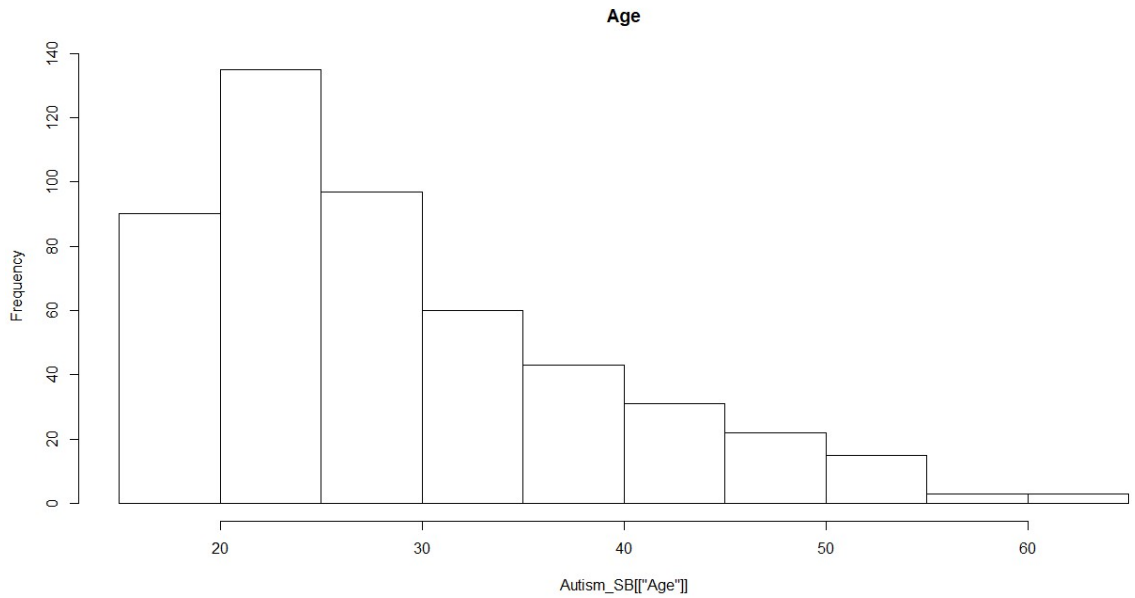
I took a slightly different approach; I sorted the ages in descending order and observed this :

```
> head(sort(Autism_SB$Age,decreasing = T),20)
 [1] 383  61  61  58  56  56  55  55  55  55  55  54  53  53  53  53  53  52  52  52
>
```

Besides 383, there were two cases 61 years old, I took the liberty of converting the outlier case into one of the 61 years old case aware of the fact that classifying the outlier into different subgroup could yield different analytical result.

Below are the histogram and boxplot after the outlier handling:





The graphical summaries above are much more aesthetic and realistic to the human age scale.

This displays the data outlier was well handled.

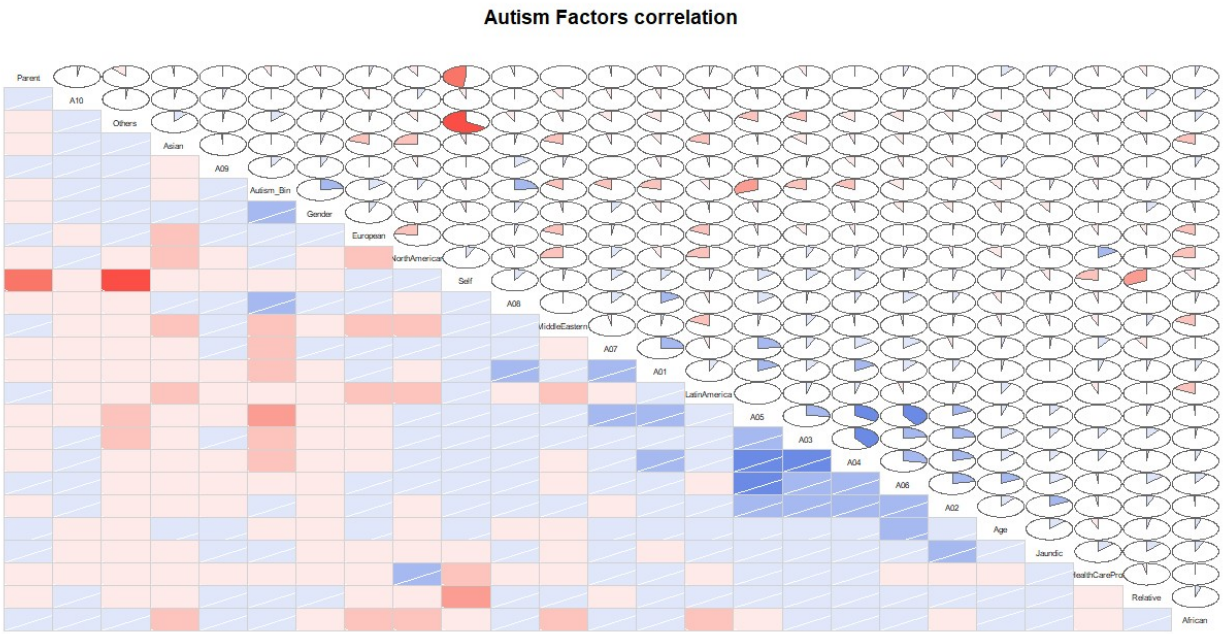
Exploratory Data Analysis

Correlations

	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	Age	Gender
A01	1.00	0.06	0.09	0.17	0.18	0.10	0.26	0.16	-0.03	-0.04	0.01	-0.06
A02	0.06	1.00	0.24	0.20	0.17	0.21	-0.03	0.05	-0.05	0.03	0.08	-0.07
A03	0.09	0.24	1.00	0.41	0.25	0.23	0.06	0.01	0.02	0.01	0.10	0.00
A04	0.17	0.20	0.41	1.00	0.33	0.28	0.14	0.05	-0.07	0.00	0.09	-0.03
A05	0.18	0.17	0.25	0.33	1.00	0.41	0.24	0.14	-0.01	-0.02	0.05	-0.04
A06	0.10	0.21	0.23	0.28	0.41	1.00	0.14	0.11	-0.04	0.03	0.13	-0.08
A07	0.26	-0.03	0.06	0.14	0.24	0.14	1.00	0.09	0.00	-0.05	0.03	0.07
A08	0.16	0.05	0.01	0.05	0.14	0.11	0.09	1.00	0.12	0.00	-0.03	0.06
A09	-0.03	-0.05	0.02	-0.07	-0.01	-0.04	0.00	0.12	1.00	0.03	0.04	0.06
A10	-0.04	0.03	0.01	0.00	-0.02	0.03	-0.05	0.00	0.03	1.00	-0.02	0.02
Age	0.01	0.08	0.10	0.09	0.05	0.13	0.03	-0.03	0.04	-0.02	1.00	-0.07
Gender	-0.06	-0.07	0.00	-0.03	-0.04	-0.08	0.07	0.06	0.06	0.02	-0.07	1.00
Jaundic	-0.01	0.18	0.07	0.09	0.09	0.14	0.02	0.01	0.00	-0.05	0.10	-0.08
Autism_Bin	-0.20	0.03	-0.19	-0.17	-0.30	-0.12	-0.16	0.23	0.08	0.01	-0.08	0.23
HealthCareProf	0.04	-0.02	0.07	0.07	0.00	-0.01	0.09	-0.03	-0.03	0.00	-0.08	-0.01
Others	-0.12	-0.06	-0.17	-0.13	-0.17	-0.12	-0.11	-0.07	0.01	0.01	-0.16	0.03
Parent	-0.03	-0.01	-0.06	0.00	-0.02	0.04	-0.02	-0.03	0.00	0.02	0.09	-0.04
Relative	0.03	0.05	0.09	0.01	0.03	0.13	-0.08	0.00	0.01	0.07	-0.01	0.08
Self	0.09	0.03	0.10	0.08	0.13	0.01	0.10	0.08	-0.01	-0.05	0.10	-0.04
African	0.05	-0.01	0.01	0.04	-0.01	0.09	0.00	0.03	0.05	0.08	0.03	-0.02
Asian	-0.07	-0.02	-0.11	-0.05	-0.01	-0.07	-0.04	0.02	-0.01	0.02	0.01	0.03
European	0.00	0.01	-0.08	-0.04	-0.03	0.01	0.02	0.02	0.01	-0.05	0.03	0.06
LatinAmerica	0.07	0.03	0.05	0.04	0.00	-0.04	-0.04	-0.03	-0.02	-0.04	0.07	-0.01
MiddleEastern	0.02	0.02	0.07	-0.02	0.02	-0.03	-0.03	0.00	0.03	-0.08	-0.03	-0.02
NorthAmerican	-0.06	-0.02	0.06	0.02	0.02	0.03	0.07	-0.03	-0.05	0.06	-0.09	-0.03

	Jaundic	Autism_Bin	HealthCareProf	Others	Parent	Relative	Self	African
A01	-0.01	-0.20	0.04	-0.12	-0.03	0.03	0.09	0.05
A02	0.18	0.03	-0.02	-0.06	-0.01	0.05	0.03	-0.01
A03	0.07	-0.19	0.07	-0.17	-0.06	0.09	0.10	0.01
A04	0.09	-0.17	0.07	-0.13	0.00	0.01	0.08	0.04
A05	0.09	-0.30	0.00	-0.17	-0.02	0.03	0.13	-0.01
A06	0.14	-0.12	-0.01	-0.12	0.04	0.13	0.01	0.09
A07	0.02	-0.16	0.09	-0.11	-0.02	-0.08	0.10	0.00
A08	0.01	0.23	-0.03	-0.07	-0.03	0.00	0.08	0.03
A09	0.00	0.08	-0.03	0.01	0.00	0.01	-0.01	0.05
A10	-0.05	0.01	0.00	0.01	0.02	0.07	-0.05	0.08
Age	0.10	-0.08	-0.08	-0.16	0.09	-0.01	0.10	0.03
Gender	-0.08	0.23	-0.01	0.03	-0.04	0.08	-0.04	-0.02
Jaundic	1.00	0.03	0.11	-0.06	0.05	0.10	-0.06	0.06
Autism_Bin	0.03	1.00	-0.04	0.10	-0.06	0.03	-0.04	0.01
HealthCareProf	0.11	-0.04	1.00	-0.05	-0.04	-0.03	-0.22	-0.01
Others	-0.06	0.10	-0.05	1.00	-0.11	-0.08	-0.66	0.06
Parent	0.05	-0.06	-0.04	-0.11	1.00	-0.06	-0.48	0.04
Relative	0.10	0.03	-0.03	-0.08	-0.06	1.00	-0.34	0.04
Self	-0.06	-0.04	-0.22	-0.66	-0.48	-0.34	1.00	-0.08
African	0.06	0.01	-0.01	0.06	0.04	0.04	-0.08	1.00
Asian	-0.01	-0.01	-0.01	0.09	-0.02	-0.03	-0.04	-0.18
European	-0.02	0.14	-0.05	0.02	0.03	-0.05	0.00	-0.17
LatinAmerica	0.00	-0.07	-0.05	-0.04	0.03	0.01	0.03	-0.17
MiddleEastern	-0.03	-0.15	-0.06	-0.04	0.00	0.06	0.02	-0.18
NorthAmerican	-0.01	0.06	0.15	-0.08	-0.07	-0.01	0.06	-0.23

	Asian	European	LatinAmerica	MiddleEastern	NorthAmerican
A01	-0.07	0.00	0.07	0.02	-0.06
A02	-0.02	0.01	0.03	0.02	-0.02
A03	-0.11	-0.08	0.05	0.07	0.06
A04	-0.05	-0.04	0.04	-0.02	0.02
A05	-0.01	-0.03	0.00	0.02	0.02
A06	-0.07	0.01	-0.04	-0.03	0.03
A07	-0.04	0.02	-0.04	-0.03	0.07
A08	0.02	0.02	-0.03	0.00	-0.03
A09	-0.01	0.01	-0.02	0.03	-0.05
A10	0.02	-0.05	-0.04	-0.08	0.06
Age	0.01	0.03	0.07	-0.03	-0.09
Gender	0.03	0.06	-0.01	-0.02	-0.03
Jaundic	-0.01	-0.02	0.00	-0.03	-0.01
Autism_Bin	-0.01	0.14	-0.07	-0.15	0.06
HealthCareProf	-0.01	-0.05	-0.05	-0.06	0.15
Others	0.09	0.02	-0.04	-0.04	-0.08
Parent	-0.02	0.03	0.03	0.00	-0.07
Relative	-0.03	-0.05	0.01	0.06	-0.01
Self	-0.04	0.00	0.03	0.02	0.06
African	-0.18	-0.17	-0.17	-0.18	-0.23
Asian	1.00	-0.18	-0.18	-0.19	-0.24
European	-0.18	1.00	-0.17	-0.18	-0.23
LatinAmerica	-0.18	-0.17	1.00	-0.18	-0.23
MiddleEastern	-0.19	-0.18	-0.18	1.00	-0.25
NorthAmerican	-0.24	-0.23	-0.23	-0.25	1.00



That is a whole lot of correlations !

The corrgram provides visual aid in cherry picking only the significant correlations.

The significant correlations are :

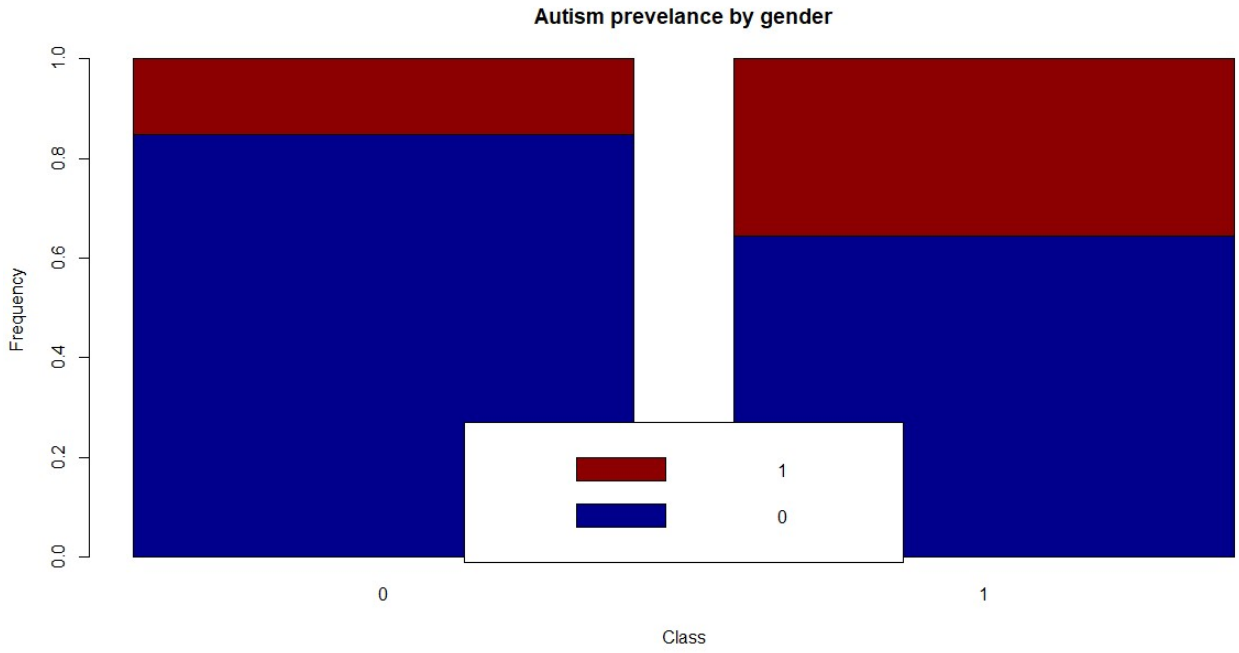
- Other and Self , -0.66
- Parent and Self, -0.48
- Relative and Self, -0.34
  If a patient himself/herself completed the test, I would guess the patient to be at the older end of the age spectrum and very unlikely to have a parent, relative or other person complete the survey for him/her. Hence, I could see “Parent”, “Other” and “Relative” variable to negatively and strongly relate against “Self” variable.

- A3 and A4, 0.41
- A4 and A5 0.33

- A5 and A6 0.41

  It would be something intrinsic between the nature of the questions above to create this correlation, since details about the questions are unknown, I could not decipher the correlation between those variables above.

  Based on the Correlation table, two most significant predictors of Autism are A05 and Gender.
  Autism_Bin and A05 have correlation of -0.30
  Autism_Bin and Gender are correlated with a factor of 0.23.
  Although the correlations above are not significant enough, but these two are the pairs with the strongest absolute value of their correlation compared to other pairs.



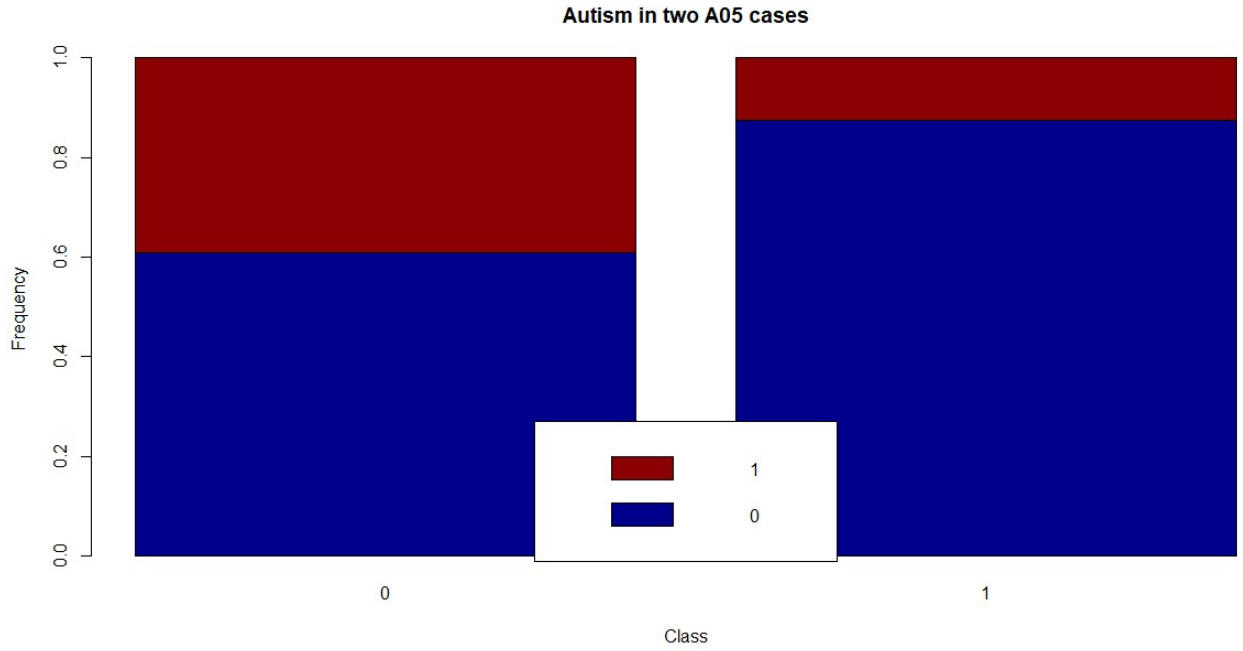
```
> summary(table(Autism_SB$Gender,Autism_SB$Autism_Bin))   #Chi-Sq
Number of cases in table: 499
Number of factors: 2
Test for independence of all factors:
        Chisq = 26.768, df = 1, p-value = 0.0000002294
> chisq.test(Autism_SB$Gender,Autism_SB$Autism_Bin)       #Chi-Sq - specific

        Pearson's Chi-squared test with Yates' continuity correction

data:  Autism_SB$Gender and Autism_SB$Autism_Bin
X-squared = 25.719, df = 1, p-value = 0.0000003948
```

The histogram by gender classification and the Chi-square test provides statistical evidence that occurrence of Autism is not the same in the two gender categories, in other words Autism prevalence depends on the case gender to some extent.

Statistical evidence for A05 :




```
> summary(table(Autism_SB$A05,Autism_SB$Autism_Bin))   #Chi-Sq
Number of cases in table: 499
Number of factors: 2
Test for independence of all factors:
        Chisq = 45.85, df = 1, p-value = 0.00000000001275
> chisq.test(Autism_SB$A05,Autism_SB$Autism_Bin)       #Chi-Sq - specific

        Pearson's Chi-squared test with Yates' continuity correction

data:  Autism_SB$A05 and Autism_SB$Autism_Bin
X-squared = 44.478, df = 1, p-value = 0.00000000002572
```

The histogram for the two A05 cases clearly displays the inconsistency of the occurrence of Autism between two A05 cases. The Chi-square test solidifies the claim by providing statistical evidence because p-value is way below 0.05.

Models

Model 1: All Variables included

1. It took 6 iterations for the model to fit the data.
2. AIC of 428 is useless right now as we do not have another model to compare its AIC
3. The residuals deviance is much lower compared to the null deviance for this model.
4. Since this is the baseline model we would expect a number of variables not contributing significantly in the model like A04, A06, A07, A09, A10 and others, any variable with z-value above 0.05 shall be removed in the forthcoming models.
5. Variable Self and NorthAmerican variables do not have any available influence in this model, perhaps this is because of their collinearity with other variables, also A05 and A08 seems to be the most significant predictors in this model.

```
Call:
glm(formula = Autism_Bin ~ A01 + A02 + A03 + A04 + A05 + A06 +
    A07 + A08 + A09 + A10 + Age + Gender + Jaundic + HealthCareProf +
    Others + Parent + Relative + Self + African + Asian + European +
    LatinAmerica + MiddleEastern + NorthAmerican, family = "binomial",
    data = Autism_SB, na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8097  -0.6093  -0.2937   0.4120   3.0295

Coefficients: (2 not defined because of singularities)
              Estimate Std. Error z value       Pr(>|z|)
(Intercept)    -1.01305    0.58525  -1.731        0.08345 .
A01            -0.85367    0.30073  -2.839        0.00453 **
A02             0.87522    0.28623   3.058        0.00223 **
A03            -0.67684    0.29945  -2.260        0.02380 *
A04            -0.35869    0.30043  -1.194        0.23251
A05            -1.69956    0.32000  -5.311 0.000000108990 ***
A06            -0.10141    0.34526  -0.294        0.76897
A07            -0.56798    0.29317  -1.937        0.05270 .
A08             2.01136    0.32400   6.208 0.000000000537 ***
A09             0.32305    0.26301   1.228        0.21934
A10            -0.18554    0.26237  -0.707        0.47948
Age            -0.01088    0.01465  -0.742        0.45780
```

```
Gender          1.41075    0.28785   4.901 0.000000953234 ***
Jaundic         0.64132    0.43704   1.467        0.14226
HealthCareProf -1.07959    1.36787  -0.789        0.42997
Others          0.13503    0.37945   0.356        0.72194
Parent         -0.87646    0.56269  -1.558        0.11932
Relative        0.61647    0.64532   0.955        0.33943
Self                 NA         NA      NA             NA
African        -0.52587    0.42575  -1.235        0.21677
Asian          -0.85948    0.41693  -2.061        0.03926 *
European        0.12968    0.39063   0.332        0.73992
LatinAmerica   -0.95343    0.44488  -2.143        0.03210 *
MiddleEastern  -1.87312    0.51283  -3.653        0.00026 ***
NorthAmerican        NA         NA      NA             NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 570.36  on 498  degrees of freedom
Residual deviance: 382.64  on 476  degrees of freedom
AIC: 428.64

Number of Fisher Scoring iterations: 6
```

Model 2: Stepwise Selection

1. Number of Fisher Scoring iterations is 5, one less than the baseline model
2. The AIC has also reduced which is important, this model is better than the baseline model.
3. The residual deviance has increased unlike the previous two factors, but it is a trade-off accepted.
4. All the variables are significantly contributing in the model except for Parent but the variable should be kept otherwise overfitting would prevail in the model.
5. The coefficients for all the variables look comparable except A08 which will strongly affect the prediction of our model.

```
Call:
glm(formula = Autism_Bin ~ A01 + A02 + A03 + A05 + A07 + A08 +
    Gender + Parent + Asian + LatinAmerica + MiddleEastern, family = "binomial",
    data = Autism_SB, na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1222  -0.6155  -0.3334   0.4665   2.8596

Coefficients:
             Estimate Std. Error z value       Pr(>|z|)
(Intercept)   -1.2207     0.3692  -3.306       0.000946 ***
A01           -0.9251     0.2930  -3.157       0.001593 **
A02            0.8634     0.2765   3.122       0.001794 **
A03           -0.8558     0.2791  -3.066       0.002168 **
A05           -1.6788     0.2900  -5.788 0.000000007112 ***
A07           -0.6515     0.2831  -2.302       0.021357 *
A08            1.9758     0.3120   6.332 0.000000000241 ***
Gender         1.3855     0.2759   5.021 0.000000512771 ***
Parent        -0.9657     0.5496  -1.757       0.078904 .
Asian         -0.8260     0.3663  -2.255       0.024118 *
LatinAmerica  -0.8673     0.3931  -2.206       0.027377 *
MiddleEastern -1.5791     0.4435  -3.560       0.000370 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 570.36  on 498  degrees of freedom
Residual deviance: 392.47  on 487  degrees of freedom
AIC: 416.47

Number of Fisher Scoring iterations: 5
```

Model 3: Forward Selection

1. It took only two iterations to fit the data.
2. The AIC is comparatively high, this model is perhaps not better than the previous two models.
3. The residual deviance changes the dynamics, a very low deviance of 65 is significantly better than previous two models.
4. All the variables look significant enough to be retained in the model.
5. The variables coefficients have decreased compared to other models, but among themselves the coefficients are in comparable range.

```
Call:
glm(formula = Autism_Bin ~ A05 + A08 + Gender + A01 + MiddleEastern +
    A03 + A02 + European + NorthAmerican + A07 + Parent + Jaundic,
    data = Autism_SB, na.action = na.omit)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.68534  -0.26625  -0.09125   0.22442   1.00185

Coefficients:
              Estimate Std. Error t value          Pr(>|t|)
(Intercept)    0.24493    0.04867   5.032 0.000000684694195 ***
A05           -0.23865    0.03575  -6.675 0.000000000067418 ***
A08            0.25905    0.03478   7.447 0.000000000000438 ***
Gender         0.17895    0.03354   5.336 0.000000146048029 ***
A01           -0.14512    0.03924  -3.699          0.000242 ***
MiddleEastern -0.09671    0.04806  -2.012          0.044757 *
A03           -0.11895    0.03521  -3.378          0.000790 ***
A02            0.09360    0.03500   2.674          0.007739 **
European       0.14073    0.04949   2.844          0.004650 **
NorthAmerican  0.09869    0.04207   2.346          0.019370 *
A07           -0.08586    0.03585  -2.395          0.017015 *
Parent        -0.10981    0.06331  -1.734          0.083477 .
Jaundic        0.09265    0.05470   1.694          0.090928 .
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1353353)

    Null deviance: 95.651  on 498  degrees of freedom
Residual deviance: 65.773  on 486  degrees of freedom
AIC: 432.93

Number of Fisher Scoring iterations: 2
```

Model 4: Model from manual iterative variable elimination

1. Number of Fischer scoring iterations is 5, which is easily acceptable.
2. The AIC is comparatively lower at 416, this model would be among the two selected models.
3. The residual is 390 comparable to the stepwise selection model.
4. All the variables look significant enough to be retained in the model.
5. African and Autism_Bin are POSITIVELY correlated but the coefficient is negative, inconsistent behaviour exhibited by the model

```
Call:
glm(formula = Autism_Bin ~ A01 + A02 + A03 + A05 + A07 + A08 +
    Gender + Parent + African + Asian + LatinAmerica + MiddleEastern,
    family = "binomial", data = Autism_SB, na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9536  -0.6234  -0.3291   0.4459   2.8657

Coefficients:
              Estimate Std. Error z value       Pr(>|z|)
(Intercept)    -1.1154     0.3785  -2.947       0.003209 **
A01            -0.8889     0.2948  -3.015       0.002568 **
A02             0.8603     0.2767   3.109       0.001876 **
A03            -0.8382     0.2800  -2.994       0.002757 **
A05            -1.7123     0.2936  -5.831 0.000000005503 ***
A07            -0.6690     0.2844  -2.352       0.018662 *
A08             1.9856     0.3126   6.351 0.000000000214 ***
Gender          1.3883     0.2769   5.013 0.000000536273 ***
Parent         -0.9436     0.5511  -1.712       0.086825 .
African        -0.5108     0.3853  -1.326       0.184880
Asian          -0.9516     0.3782  -2.516       0.011862 *
LatinAmerica   -1.0035     0.4062  -2.470       0.013496 *
MiddleEastern  -1.7116     0.4548  -3.763       0.000168 ***
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 570.36  on 498  degrees of freedom
Residual deviance: 390.66  on 486  degrees of freedom
AIC: 416.66

Number of Fisher Scoring iterations: 5
```

Model Evaluation

Stepwise Selection model

The confusion matrix has been created in R :

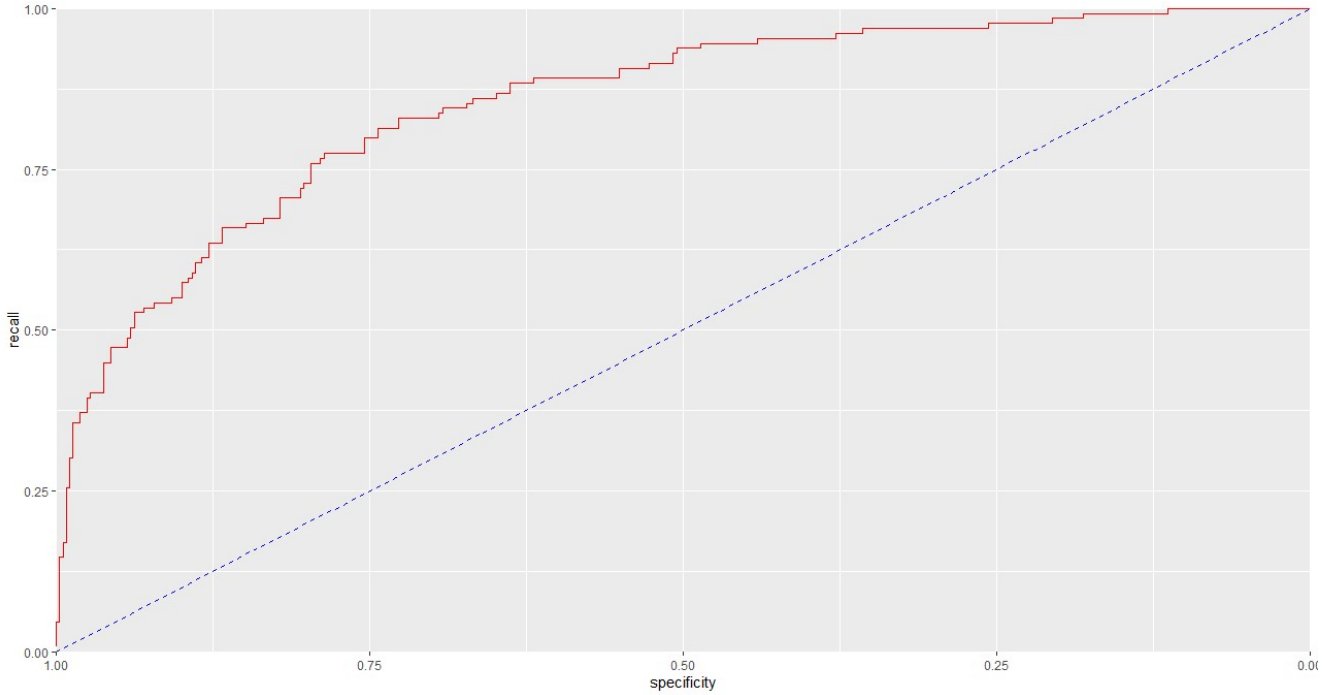
```
       Yhat = 1 Yhat = 0
Y = 1        85       44
Y = 0        53      317
```

a. Accuracy
   0.805, high accuracy; so, 80.5% of the cases were correctly classified
b. Specificity
   0.856 ; 85.6% of the 0 class datapoints were classified as 0
c. Sensitivity
   0.658 ; only 65.8% of the 1 class datapoints were classified as 1

d. Precision

0.615 ; 61.5% of the predicted class 1 datapoints were actually class 1 datapoints rest of the 28.5% were incorrectly predicted as class 1 datapoints

ROC curve

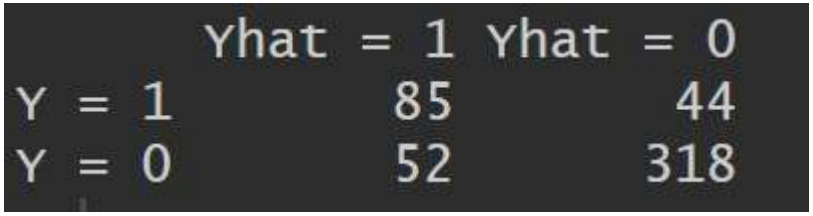


Area Under Curve : 0.853

The ROC curve displays the trade-offs between the sensitivity and specificity, our model is significantly better than a standard random classifier.

The AUC highlights the general predictive capability of the model, the closer it is to one the better the model. Our model has a decent AUC of 0.853

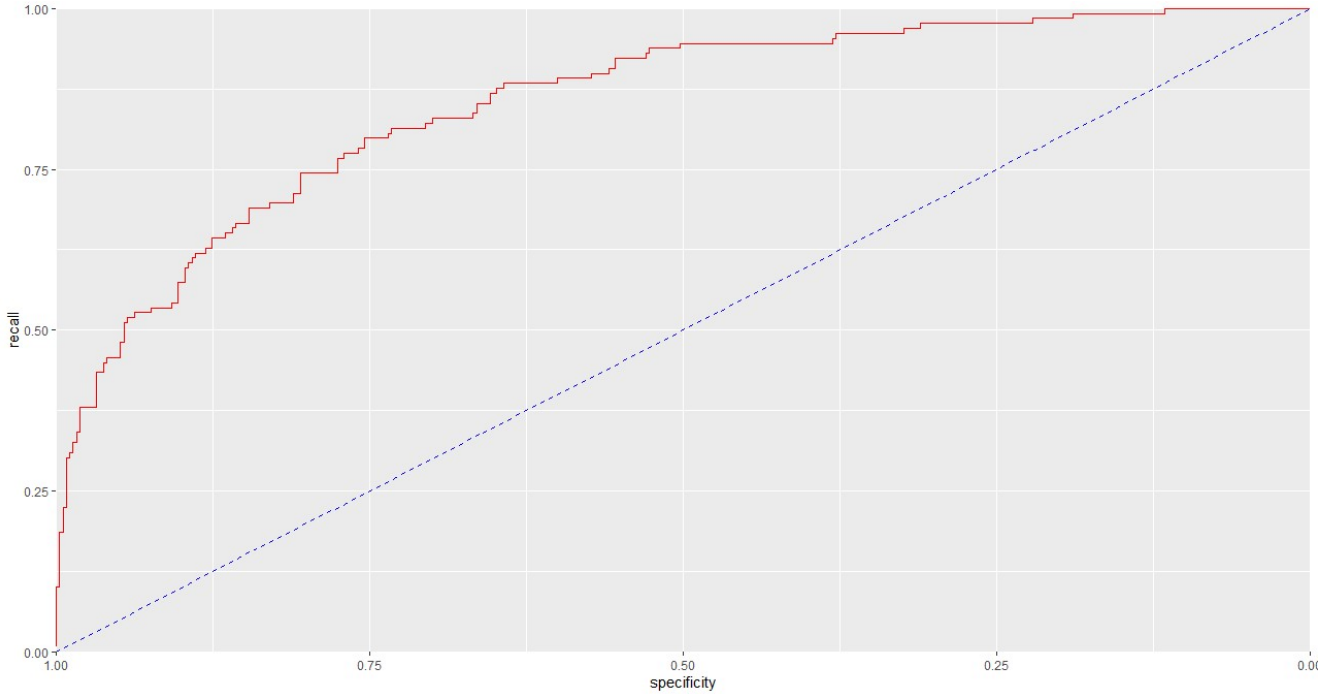
Model from manual iterative elimination



	Yhat = 1	Yhat = 0
Y = 1	85	44
Y = 0	52	318

e. Accuracy
   0.807, high accuracy; so, 80.7% of the cases were correctly classified
f. Specificity
   0.859 ; 85.9% of the 0 class datapoints were classified as 0
g. Sensitivity
   0.658 ; only 65.8% of the 1 class datapoints were classified as 1
h. Precision
   0.62 ; 62% of the predicted class 1 datapoints were actually class 1 datapoints rest of the 28.5% were incorrectly predicted as class 1 datapoints

ROC curve



Area Under Curve : 0.854

The AUC for this model is approximately the same as the previous model.

The ROC curve for this model is decent, it is well in the top left section of the graph which is the preferred trait of the ROC curve.

Final Model, Recommendation, and Interpretation

Two models selected for comparison :

Stepwise selection model (model number 2)

Iterative Model (model number 4)

The AUC for both the model are approximately the same.

The AIC is also the same for both the models. The residual deviance for the iterative model is slightly smaller, which adds weight in its favor. Although the number of predictors in the stepwise model is one less than the iterative model, I would still recommend the Iterative model because I think the iterative model is less prone to overfitting than the Stepwise selection model.

Autism_Bin = 0.86*A02 - 0.88*A01 - 0.83*A03 – 1.7*A05 – 0.66*A07 + 1.9*A08 +

1.3*Gender – 0.94*Parent - 0.51*African – 0.95Asian – LatinAmerica – 1.7*MiddleEastern

APPENDIX 1: Data Transformation

Autism Binary variable creation

```
Autism_SB$Autism_Bin<-as.numeric(Autism_SB$Autism)
Autism_SB$Autism_Bin<-Autism_SB$Autism_Bin-1
```

Nationality Dummy variable creation

```
Nationality_Dummies_SB <- model.matrix(~Nationality -1, data=Autism_SB)
```

Rel Dummy variable creation

```
Rel_Dummies_SB<-model.matrix(~Rel -1, data=Autism_SB)
```

Age Outlier handling

```
Autism_SB[Autism_SB$Age>61,'Age']<-61
```