Predicting the Age and Gender of Abalone

Introduction:

Abalone is an edible mollusk of warm seas, with a shallow ear-shaped shell lined with mother-of-pearl and pierced with a line of respiratory holes.

Data Dictionary

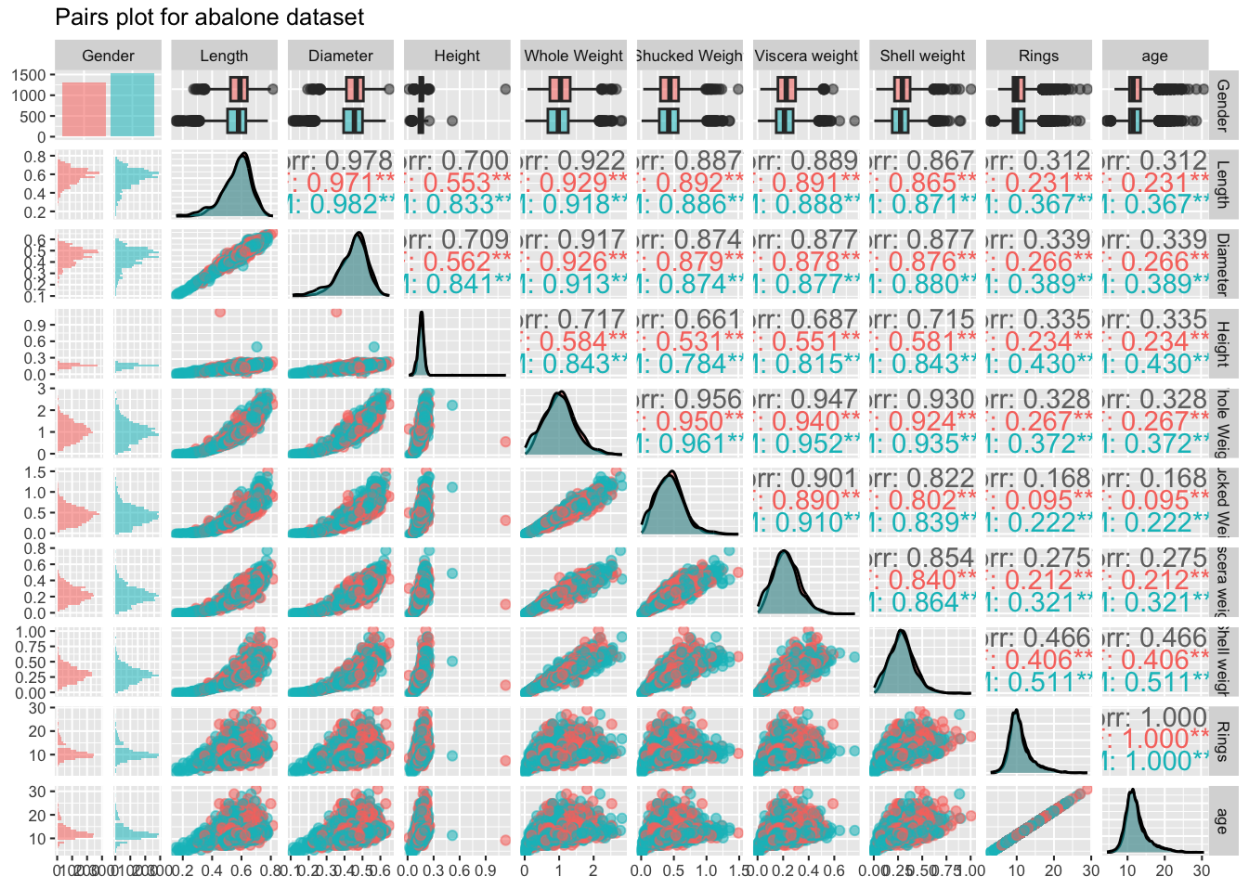
Name	Data Type	Unit	Description
Gender	nominal		M, F
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to the length
Height	continuous	mm	with meat in the shell
Whole weight	continuous	grams	of the whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut-weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer		+1.5 gives the age in years

Questions:

1. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. You will build a regression model to determine the age of the Abalone from the other attributes. How good is your model? Make sure your residual analysis shows no patterns and is normally distributed. What attributes proved valuable in predicting age?
2. Once you have the model, use the rings column to determine the accuracy measure of your model. Remember, just pick one measure of accuracy (MAPE would be my recommendation)
3. Are you able to predict the gender from the other attributes? What is the accuracy of using logistic regression? What is the accuracy of using LDA? What about the other measures of accuracy?
4. Do dimension reduction techniques make your classification model better?

Working:

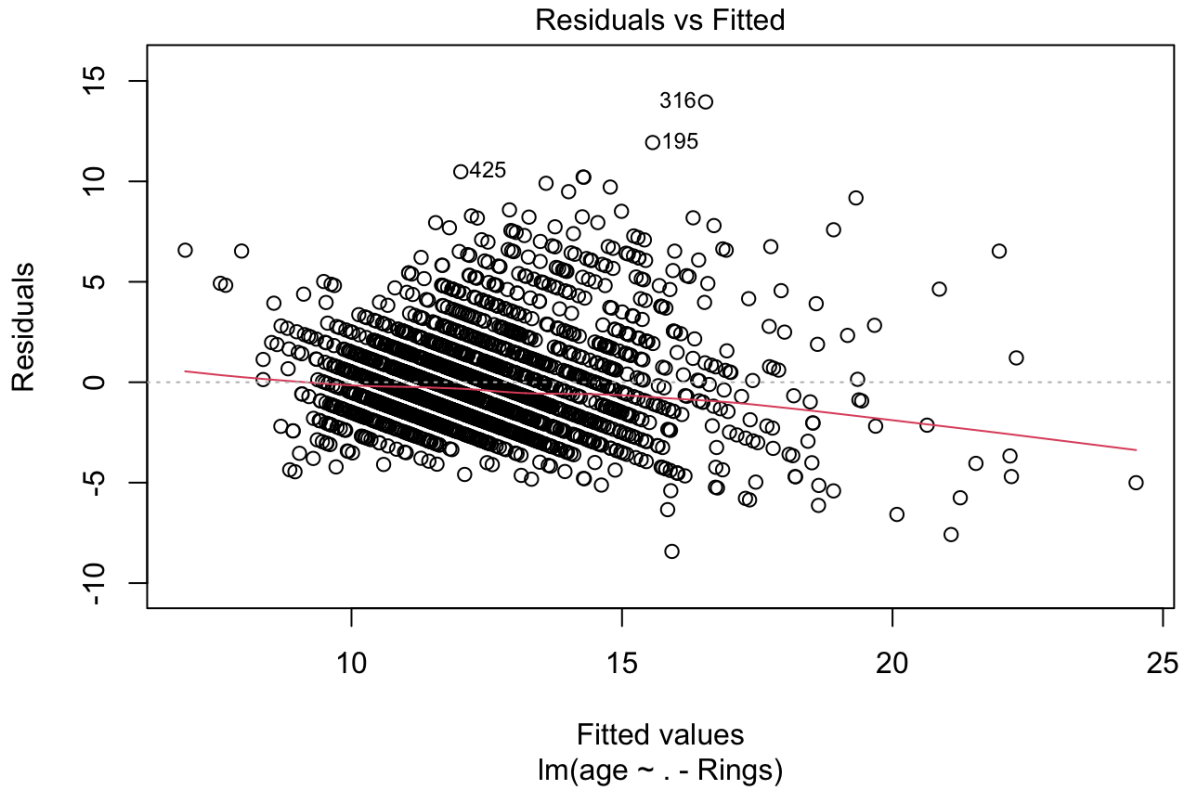
The pair plot for our data set is as follows:

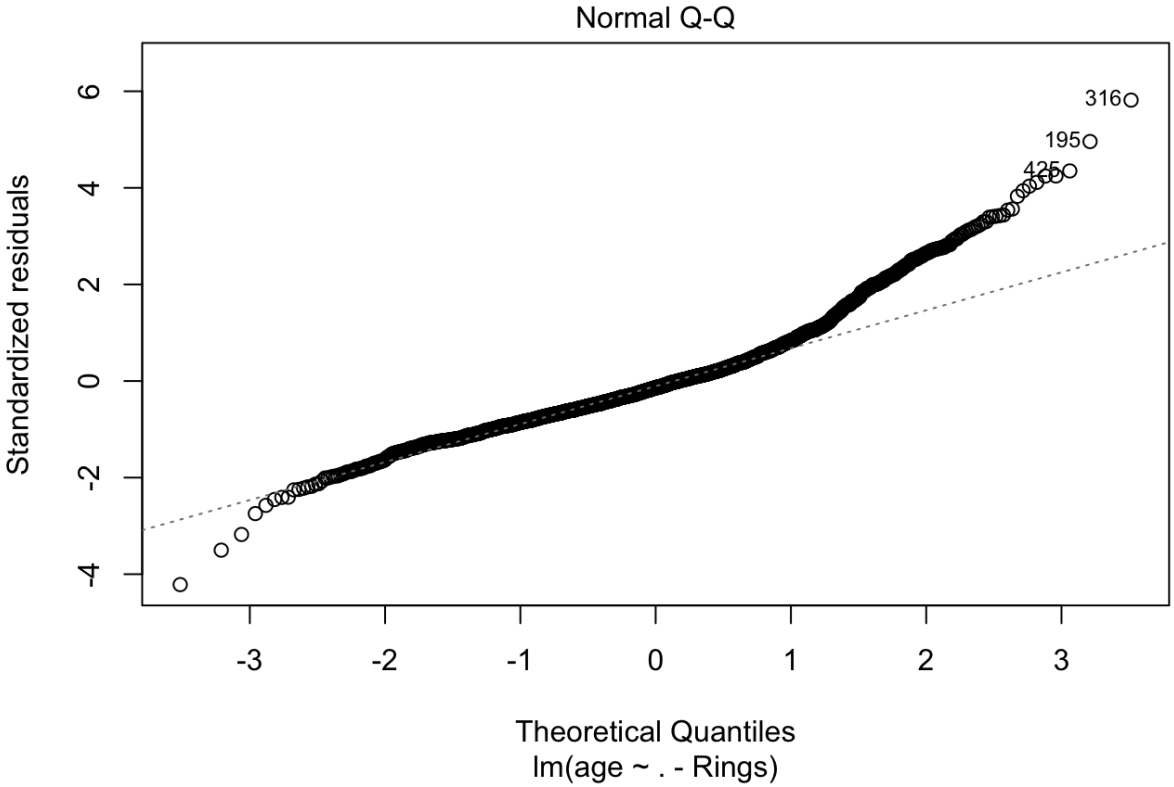


Observations:

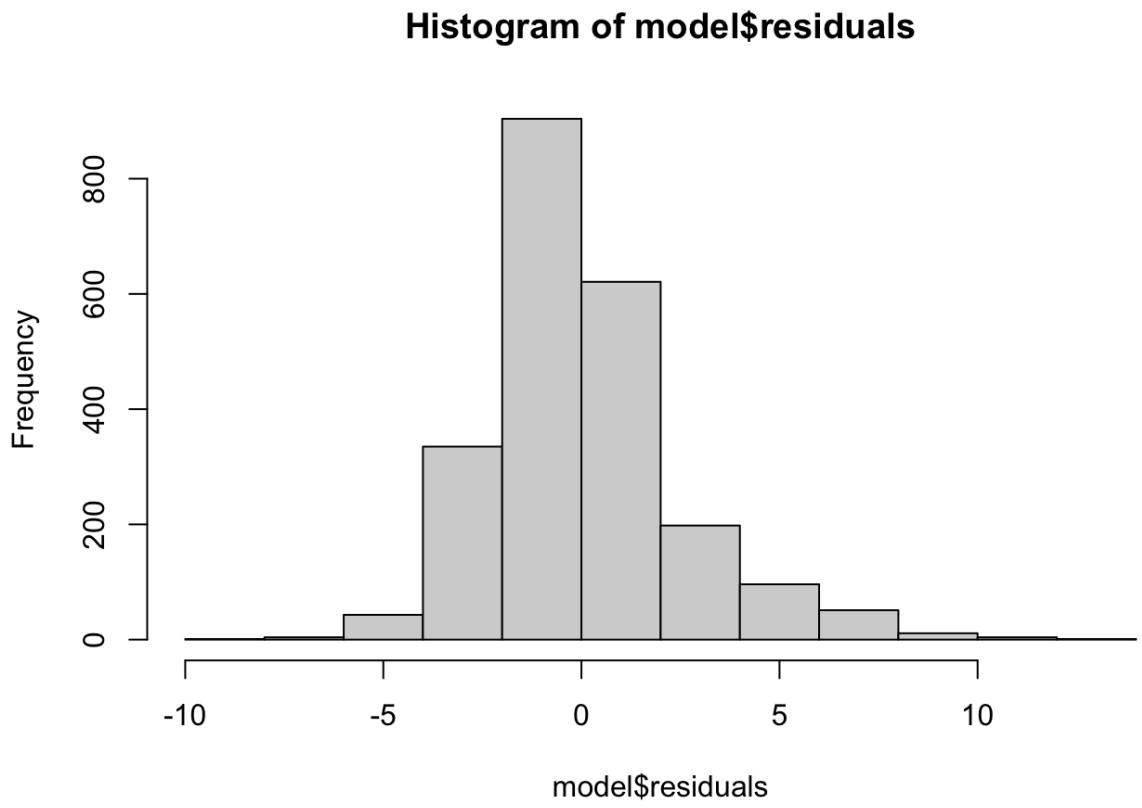
- The data have a strong correlation, which is the first thing to remark. The predictors appear to be very multicollinear with one. 'Diameter' and 'Length,' for instance, have a connection that is exceptionally high (around 97.8).
- Similar to "Shucked_weight," "Viscera_weight," and "Shell_weight," "Whole_weight" also appears to have a strong correlation with other weight predictors.
- Second, compared to all other predictors, the distributions of the predictor "Sex" with factor level values of "female" and "male" are highly comparable.

1. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. You will build a regression model to determine the age of the Abalone from the other attributes. How good is your model? Make sure your residual analysis shows no patterns and is normally distributed. What attributes proved valuable in predicting age?
2. Once you have the model, use the rings column to determine the accuracy measure of your model. Remember, just pick one measure of accuracy (MAPE would be my recommendation)




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(age ~ . - Rings)
316
195
425
-4
-2
0
2
4
6
-3
-2
-1
0
1
2
3

```
##
## Call:
## lm(formula = age ~ . - Rings, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4231 -1.5343 -0.3396  1.0193 13.9546
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.04870    0.56673  14.202  < 2e-16 ***
## GenderM           -0.03153    0.10291  -0.306 0.759372
## Length            -4.23017    2.68044  -1.578 0.114668
## Diameter           9.59755    3.16881   3.029 0.002483 **
## Height             6.30711    1.86644   3.379 0.000739 ***
## `Whole Weight`    10.28429    0.96566  10.650  < 2e-16 ***
## `Shucked Weight` -20.55490    1.10168 -18.658  < 2e-16 ***
## `Viscera weight` -10.33361    1.66908  -6.191 7.07e-10 ***
## `Shell weight`     8.23439    1.51571   5.433 6.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.411 on 2260 degrees of freedom
## Multiple R-squared:  0.3877, Adjusted R-squared:  0.3855
## F-statistic: 178.8 on 8 and 2260 DF,  p-value: < 2.2e-16
```



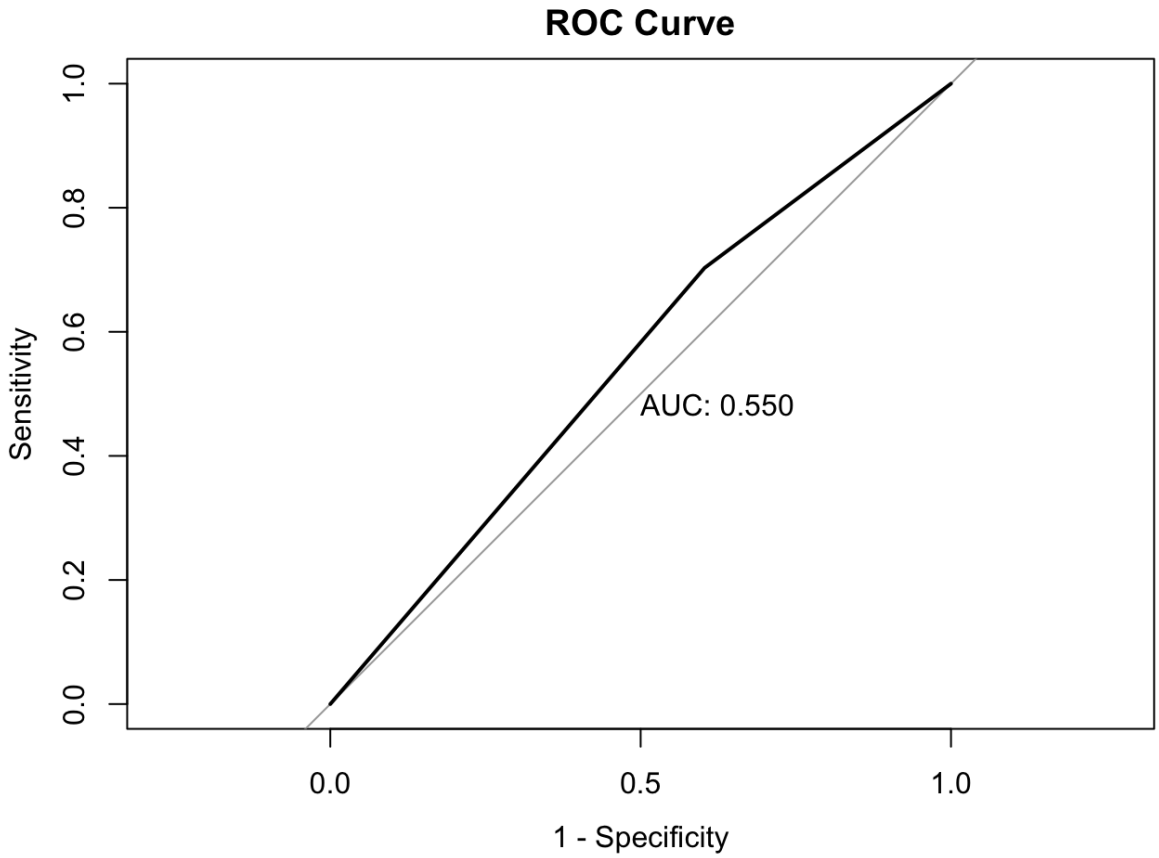
Observations:

- The p-values associated with most of the predictor variables are less than 0.05, indicating that these variables are statistically significant in predicting the age of the abalone. These variables are diameter, height, whole weight, shucked weight, viscera weight, and shell weight.
- The p-value for the Gender variable is greater than 0.05, indicating that it is not statistically significant in predicting the age of the abalone.
- The Adjusted R-squared value is 0.3855, indicating that the predictor variables explain 38.55% of the variation in the response variable, age.
- The F-statistic value is 178.8 with a p-value of < 2.2e-16, indicating that the model is significant.
- RMSE value is 2.291639
- MAPE value is 14.3003
- The Histogram of the residuals in the model shows a normal distribution.
- QQ plot shows that the data have more extreme values than would be expected.

3. Are you able to predict gender from the other attributes? What is the accuracy of using logistic regression? What is the accuracy of using LDA? What about the other measures of accuracy?

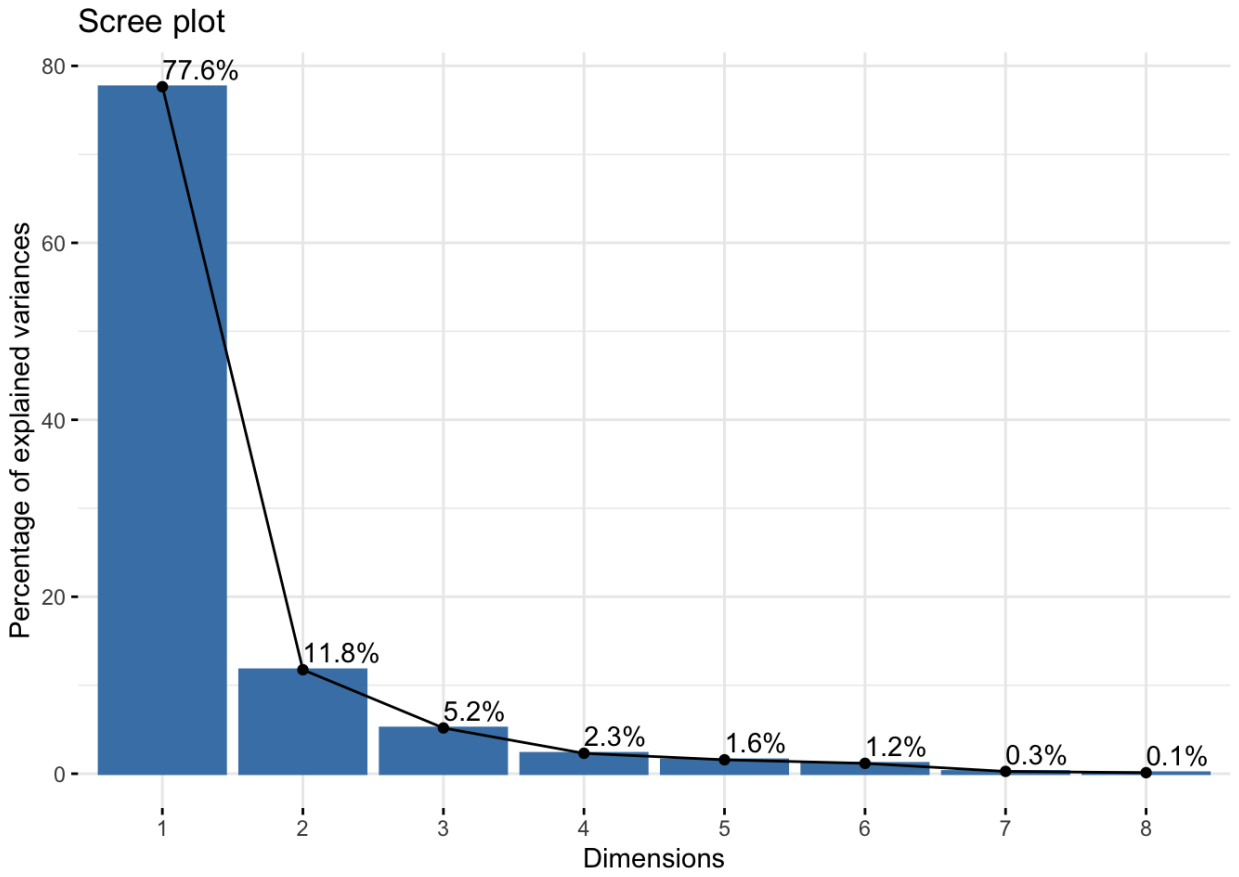
>

```
##
## lda_pred   0   1
##        0 155 137
##        1 238 321
```



Observation:
The model has a weak capacity to differentiate between positive and negative classes, as indicated by an AUC (Area Under the Curve) of 0.55. The AUC of a random guessing model would be 0.5, indicating that the model's performance is only marginally superior to random. As confirmed by the confusion matrix, the false positive rate is very high.

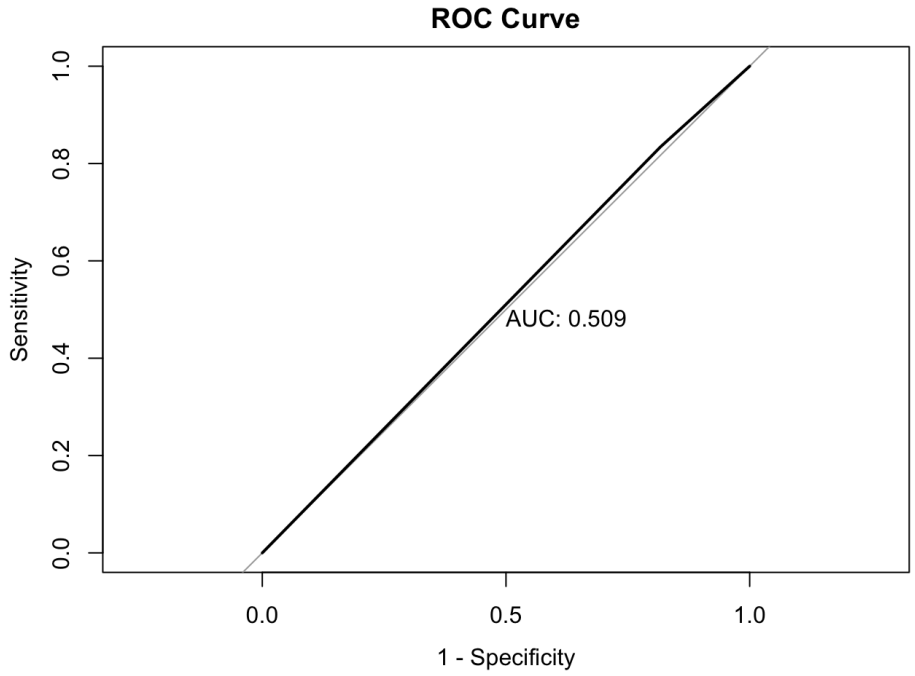
4. Do dimension reduction techniques make your classification model better?
Doing PCA first as a dimension reduction technique.




```
summary(mydata_pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.4923 0.9697 0.6431 0.42965 0.35503 0.30635 0.14437
## Proportion of Variance 0.7764 0.1175 0.0517 0.02307 0.01576 0.01173 0.00261
## Cumulative Proportion  0.7764 0.8940 0.9457 0.96874 0.98450 0.99623 0.99884
##                           PC8
## Standard deviation     0.09646
## Proportion of Variance 0.00116
## Cumulative Proportion  1.00000
```

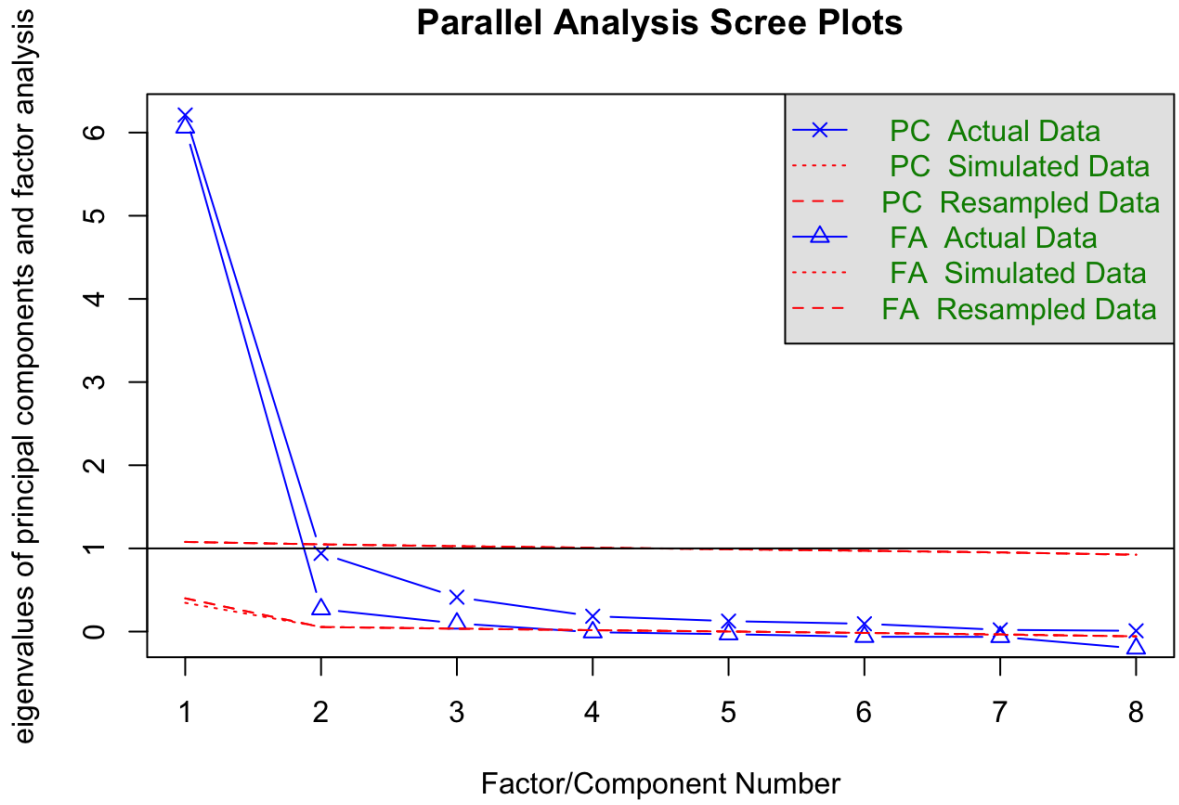
```
##
## Call:
## glm(formula = Gender ~ . - Gender, family = binomial, data = train_data_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4716  -1.2353   0.9901   1.1095   1.4959
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.15618    0.04530   3.448 0.000565 ***
## PC1         -0.06415    0.01794  -3.576 0.000348 ***
## PC2          0.10211    0.04471   2.284 0.022381 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2738.1  on 1983  degrees of freedom
## Residual deviance: 2719.5  on 1981  degrees of freedom
## AIC: 2725.5
##
## Number of Fisher Scoring iterations: 4
```



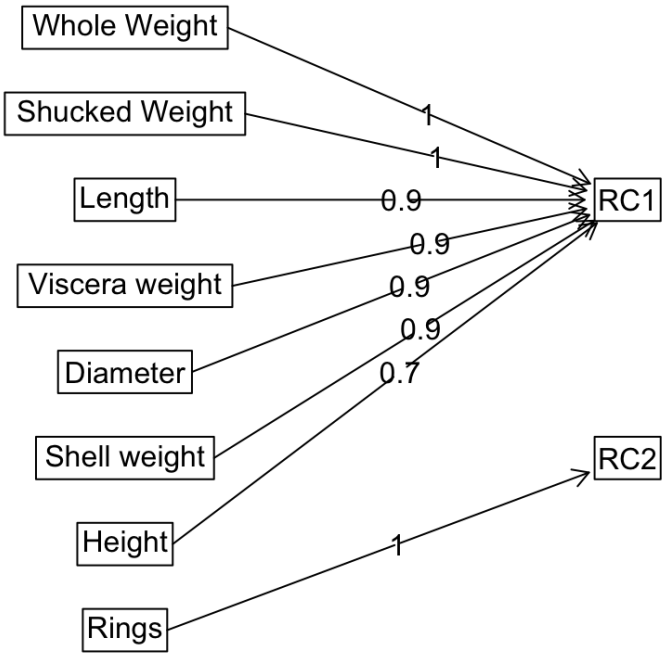
Observations:

After performing logistic regression on PCs, the result was not improved as AUC comes to be 50.9%

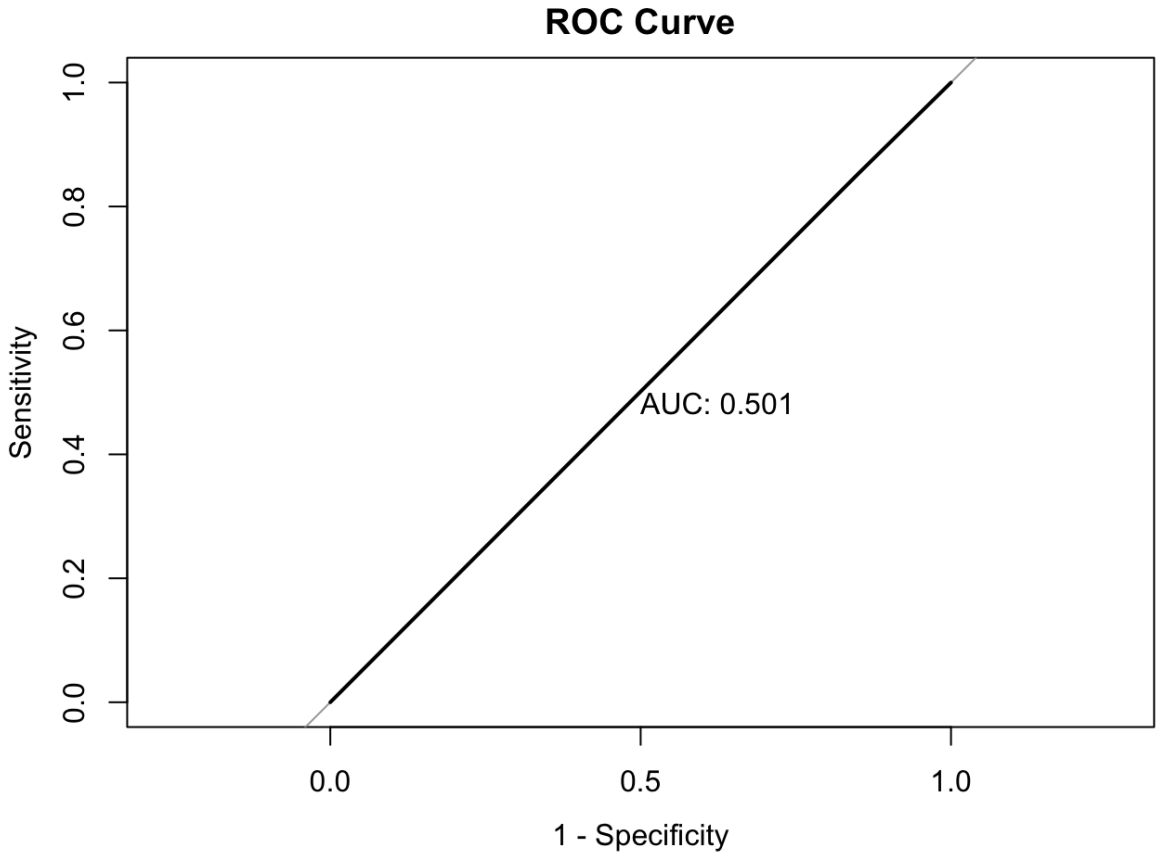
Let’s see if we get better results with EFA.



Components Analysis



```
##
## Call:
## glm(formula = Gender ~ . - Gender, family = binomial, data = train_data_2
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4671  -1.2331   0.9944   1.1105   1.4796
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.15764    0.04235   3.723 0.000197 ***
## RC1         -0.12843    0.04243  -3.027 0.002471 **
## RC2         -0.13052    0.04095  -3.187 0.001435 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3130.4  on 2267  degrees of freedom
## Residual deviance: 3110.6  on 2265  degrees of freedom
## AIC: 3116.6
##
## Number of Fisher Scoring iterations: 4
```



Observations:
Even after performing logistic regression on Factors, the result was not improved as AUC comes to be 50.1% only, so we can conclude that in this case, dimension reduction techniques did not make the classification model better.