STA 3000 Project

Adrian Harris

12/17/2019

Data set

https://www.kaggle.com/johnsmith88/heart-disease-dataset

Introduction

Principal Component Analysis

I wanted to practice using different learning algorithms on an unlabeled heart disease dataset. This challenged my already excessing knowledge. Using Principal Component Analysis, I wanted to find the values that contribute to classifying the age of the person in this data set. After running the dimension reduction method for the data set I visualized the data. I found out the variance in the data set was mostly on the first PC and the second PC. I also found out the variables that contribute to the most to the classification of age on the newly transformed variables of PC1 and PC2. After doing this I conducted a multi logistic regression on the predictions of the training and the test set for the gender. Then I created a confusion matrix to get the error of misclassification. Overall PCA was a good way to understand the correlation between variables in this data set when they are transformed into "two new variables" on PC1 and PC2.

Neural Network

First I normalized the data set using the min-max method to get scale the data variables on an interval of 0-1. I didn't scale the age variable at first because of the fear of losing significance in the variable. I tried different ways to predict sex. I used one hidden layer with one node. I also moved on with two hidden layers with multiple nodes. I used the same method of backpropagation for calculating the weights at each node. My overall experience with this deep learning algorithm was not great. The model wasn't allowing me to calculate the weights of a neural network that used more than 5 nodes in each layer.

Logistic Regression

I also performed a logistics regression on this data set to classify sex. I used backward selection to see what variables to use when it comes to predicting gender in this data set. I ended leaving all the variables in the model. You can see similarities between the PCA plot and the logistic regression model. Such variables like chol and thal contribute to predicting gender or age.

Analysis

PCA

Libararies

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
##
## Attaching package: 'neuralnet'
```

```
## The following object is masked from 'package:dplyr':
##
##     compute
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

Data set

```
## Parsed with column specification:
## cols(
##   age = col_double(),
##   sex = col_double(),
##   cp = col_double(),
##   trestbps = col_double(),
##   chol = col_double(),
##   fbs = col_double(),
##   restecg = col_double(),
##   thalach = col_double(),
##   exang = col_double(),
##   oldpeak = col_double(),
##   slope = col_double(),
##   ca = col_double(),
##   thal = col_double(),
##   target = col_double()
## )
```

Pre Processing

Partitioning of the data and setting the interations of the sampling.

Model

Removing the response variable (Gender). Scaling the data to remove noise. Center for getting the averages.

Attributes in model.

```
## $names
## [1] "sdev"     "rotation" "center"   "scale"    "x"
##
## $class
## [1] "prcomp"
```

Averages of variables in the data set

```
##         age          cp    trestbps        chol         fbs     restecg
##  54.2979943   0.9484241 131.7836676 246.6504298   0.1418338   0.5229226
##     thalach       exang     oldpeak       slope          ca        thal
## 149.2722063   0.3452722   1.1060172   1.3724928   0.7679083   2.3166189
##      target
##   0.5071633
```

PC11-13 don't play much imporantance in the variablity. PC1 captures the most of the variance in the data set.

```
## Importance of components:
##                           PC1    PC2     PC3     PC4    PC5     PC6
## Standard deviation     1.8300 1.2428 1.09109 1.00213 0.9861 0.96124
## Proportion of Variance 0.2576 0.1188 0.09158 0.07725 0.0748 0.07108
## Cumulative Proportion  0.2576 0.3764 0.46799 0.54524 0.6200 0.69112
##                            PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.93410 0.86869 0.83691 0.71799 0.66621 0.61943
## Proportion of Variance 0.06712 0.05805 0.05388 0.03965 0.03414 0.02951
## Cumulative Proportion  0.75824 0.81628 0.87016 0.90982 0.94396 0.97347
##                           PC13
## Standard deviation     0.58724
## Proportion of Variance 0.02653
## Cumulative Proportion  1.00000
```

Correlation between variables on the new Principal Components in the model.

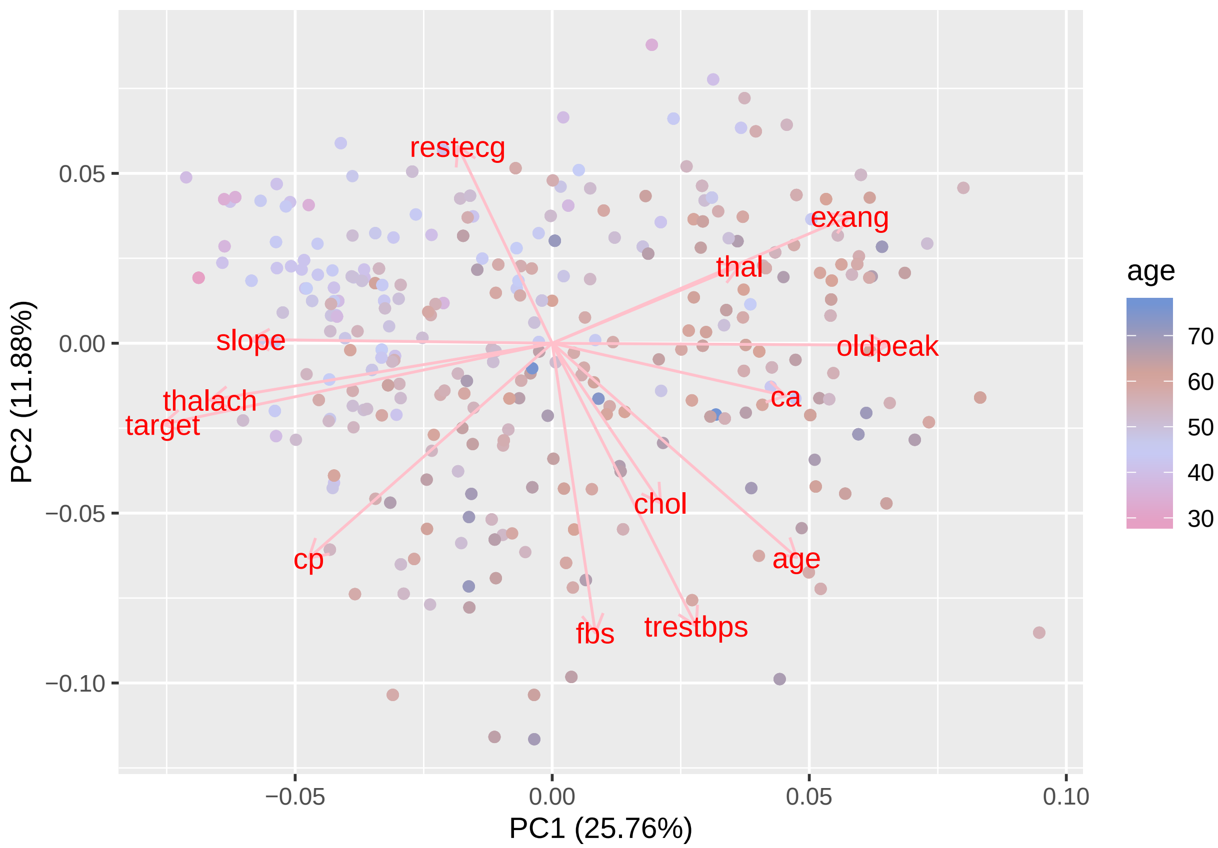
```
## Standard deviations (1, .., p=13):
##  [1] 1.8299505 1.2428420 1.0910913 1.0021266 0.9861170 0.9612405 0.9341003
##  [8] 0.8686927 0.8369062 0.7179863 0.6662148 0.6194291 0.5872440
##
## Rotation (n x k) = (13 x 13):
##                  PC1          PC2         PC3         PC4         PC5
## age       0.27113348 -0.359731809  0.06831842 -0.03453724 -0.17111184
## cp       -0.27007155 -0.360698052 -0.30506447  0.07309110 -0.34334242
## trestbps  0.15982626 -0.474652547 -0.14200047  0.21268858 -0.03284804
## chol      0.12019718 -0.268163331  0.33269257  0.59696926  0.12993091
## fbs       0.04783735 -0.486077250  0.04078210 -0.45677758  0.10998554
## restecg  -0.10474741  0.330581562 -0.24379058 -0.01403844 -0.49689753
## thalach  -0.37956484 -0.095469767  0.10365939  0.16387199 -0.03981616
```

```
## exang     0.33029474  0.213115137  0.04889392 -0.01225543  0.38804289
## oldpeak   0.37207068 -0.003304975 -0.41506603  0.11384276 -0.15164128
## slope    -0.33402156  0.006414684  0.54445300 -0.10564437 -0.09669798
## ca        0.25921481 -0.088384244  0.32987429 -0.43925070 -0.40530844
## thal      0.20776977  0.129663956  0.31355118  0.36627834 -0.47139835
## target   -0.43216211 -0.136084315 -0.14610477  0.06918944  0.06980103
##                  PC6         PC7         PC8         PC9        PC10
## age      -0.54940646 -0.01447439  0.34908436 -0.10455745 -0.13503632
## cp        0.03086459 -0.10152229 -0.04908567 -0.38211039  0.60323103
## trestbps  0.24651911  0.28252699  0.36642929  0.58437810  0.10928942
## chol     -0.29975527  0.10820415 -0.55473065 -0.01552946  0.03146873
## fbs       0.24848937  0.43243988 -0.23802814 -0.34501022 -0.32438750
## restecg  -0.25652832  0.64619181 -0.21029434  0.15369037 -0.06245594
## thalach   0.43445081 -0.04563677 -0.22863642  0.22550560 -0.11540284
## exang     0.14472024  0.40382909  0.02711903 -0.15104526  0.57980734
## oldpeak   0.14809540 -0.13413648 -0.25493611  0.08467181 -0.09402544
## slope    -0.07412110  0.20271873  0.18852406  0.14458366  0.17754863
## ca        0.03840624 -0.25084708 -0.32482439  0.27397643  0.28145286
## thal      0.40233935  0.06840515  0.26777340 -0.41890950 -0.16270276
## target   -0.13620442  0.03917052  0.04578006 -0.07986390 -0.01638585
##                  PC11         PC12        PC13
## age       0.482356541  0.165691171  0.22075971
## cp       -0.097737463  0.216096403 -0.02919813
## trestbps -0.209677834 -0.101276275  0.01239997
## chol     -0.129890108 -0.034177944 -0.03382173
## fbs      -0.052581848  0.009650356 -0.08067744
## restecg  -0.014471568  0.041182167  0.12721836
## thalach   0.537410737  0.272518134  0.37051257
## exang     0.350316344 -0.074459036  0.14267093
## oldpeak   0.423054402 -0.024702595 -0.59397375
## slope     0.173319376  0.147169837 -0.62063797
## ca        0.008830707 -0.324440489  0.17156389
## thal     -0.018354034 -0.209095845  0.03247004
## target    0.270199404 -0.813371071 -0.01738548
```

Visualzation of PCA

This plot visualizes the correlation between variables in the training set. PC1 and PC2 transform all the original variables into two variables. For example, on PC1, cholesterol is given a positive value on this new variable. When it comes to looking at age, such variables as resting blood pressure and cholestorol are highly correlated.

```
## Scale for 'colour' is already present. Adding another scale for
## 'colour', which will replace the existing scale.
```



Prediction

I used multidimensional logistic regression to predict gender based on the most important PC's (1-2)

Prediction on training

Prediction on Test

Multidimensional Logistic Regression on PC1 and PC2

```
## # weights:  4 (3 variable)
## initial  value 483.816732
## final  value 408.309332
## converged
```

```
## Call:
## multinom(formula = sex ~ PC1 + PC2, data = pcatrain)
##
## Coefficients:
##                 Values  Std. Err.
## (Intercept) 0.9091983 0.08634390
## PC1         0.1634229 0.04781332
## PC2         0.3054156 0.06778321
##
## Residual Deviance: 816.6187
## AIC: 822.6187
```

Confusion Matrix

Training set

Males = 1 Females = 0

Male count

```
## [1] 491
```

Female count

```
## [1] 207  14
```

Females are getting classifed as males on the training set

```
##
## p     0   1
##   0   9  15
##   1 198 476
```

```
## [1] 0.3051576
```

Test set

Matrix Same occurrence on the training set.

```
##
## p1    0   1
##   0   0   5
##   1 105 217
```

Error

```
## [1] 0.3363914
```

Conclusion

The model does not seem to be overfitting because the errors seem to be around the same but they are still high. When you add all the Principal Components to the model the error goes down to the error of the logistic regression further in the analysis.

Nueral Network

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 1025 obs. of  14 variables:
##  $ age     : num  52 53 70 61 62 58 58 55 46 54 ...
##  $ sex     : num  1 1 1 1 0 0 1 1 1 1 ...
##  $ cp      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ trestbps: num  125 140 145 148 138 100 114 160 120 122 ...
##  $ chol    : num  212 203 174 203 294 248 318 289 249 286 ...
##  $ fbs     : num  0 1 0 0 1 0 0 0 0 0 ...
```

```
##  $ restecg : num  1 0 1 1 1 0 2 0 0 0 ...
##  $ thalach : num  168 155 125 161 106 122 140 145 144 116 ...
##  $ exang   : num  0 1 1 0 0 0 0 1 0 1 ...
##  $ oldpeak : num  1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...
##  $ slope   : num  2 0 0 2 1 1 0 1 2 1 ...
##  $ ca      : num  2 0 0 1 3 0 3 1 0 2 ...
##  $ thal    : num  3 3 3 3 2 2 1 3 3 2 ...
##  $ target  : num  0 0 0 0 0 1 0 0 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   age = col_double(),
##   ..   sex = col_double(),
##   ..   cp = col_double(),
##   ..   trestbps = col_double(),
##   ..   chol = col_double(),
##   ..   fbs = col_double(),
##   ..   restecg = col_double(),
##   ..   thalach = col_double(),
##   ..   exang = col_double(),
##   ..   oldpeak = col_double(),
##   ..   slope = col_double(),
##   ..   ca = col_double(),
##   ..   thal = col_double(),
##   ..   target = col_double()
##   .. )
```

Normalization

You normalize the data by using a few ways. I choose the minimum-maximum transformation.

Neural Network model one node in the hidden layer

Prediction

Got out the response varaible in the prediction

Found the probabiltiies of a male or female

```
##           [,1]
## [1,] 0.9966585
## [2,] 0.9075560
## [3,] 0.6609205
## [4,] 0.9715471
## [5,] 0.4930223
## [6,] 0.4614049
```

```
## # A tibble: 1 x 14
##     age   sex    cp trestbps  chol   fbs restecg thalach exang oldpeak
##   <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl>   <dbl>   <dbl> <dbl>   <dbl>
## 1 0.479     1     0    0.292 0.196     0     0.5   0.740     0   0.161
## # ... with 4 more variables: slope <dbl>, ca <dbl>, thal <dbl>,
## #   target <dbl>
```

Confusion matrix - Training set

```
##
## predictt   0   1
##        0  91  71
##        1 116 420
```

Error

```
## [1] 0.2679083
```

Confusion matrix - Test set

```
##
## predictt2   0   1
##         0  42  35
##         1  63 187
```

Error

```
## [1] 0.2996942
```

Network with 5 nodes in hidden layer

Some weights weren't calculated when it was run so this means that a confusion matrix could not be reached.

```
## Warning: Algorithm did not converge in 1 of 1 repetition(s) within the
## stepmax.
```

Network with two hidden layers

Some weights weren't calculated when it was run so this means that a confusion matrix could not be reached.

```
## Warning: Algorithm did not converge in 1 of 1 repetition(s) within the
## stepmax.
```

Conclusion

Even though data were normalized to remove the noise of each variable the network couldn't handle more than one node in the hidden layer to make an accurate calculation for the weights. This may be from just having too much data in the model or the data wasn't scaled correctly to handle this model. Further analysis will have to be done using this model.

Logistic Regression

Partionationg of the data

Logistic Regression model

More stars are the better predictors of sex

```
##
## Call:
## glm(formula = sex ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1371  -0.9225   0.4891   0.7909   1.7646
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.28902    0.79483   2.880 0.003978 **
## age         -1.67644    0.59144  -2.835 0.004589 **
## cp           0.85205    0.32864   2.593 0.009523 **
## trestbps    -1.64174    0.61206  -2.682 0.007311 **
## chol        -4.53013    0.87565  -5.173 2.30e-07 ***
## fbs          0.27532    0.28077   0.981 0.326810
## restecg     -0.74370    0.36531  -2.036 0.041771 *
## thalach      0.10116    0.69117   0.146 0.883634
## exang        0.40419    0.24587   1.644 0.100186
## oldpeak      0.43592    0.69193   0.630 0.528691
## slope        1.18695    0.39571   3.000 0.002704 **
## ca          -0.06109    0.41936  -0.146 0.884182
## thal         1.65410    0.50044   3.305 0.000949 ***
## target      -1.87995    0.29147  -6.450 1.12e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 848.66  on 697  degrees of freedom
## Residual deviance: 706.47  on 684  degrees of freedom
## AIC: 734.47
##
## Number of Fisher Scoring iterations: 4
```

Prediction

Probabiltiies assocatied with the obervastions

```
##         1         2         3         4         5         6
## 0.9370259 0.9281351 0.8148236 0.8871616 0.6063224 0.4619173
```

```
## # A tibble: 6 x 14
##     age   sex    cp trestbps  chol   fbs restecg thalach exang oldpeak
##   <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl>   <dbl>   <dbl> <dbl>   <dbl>
```

```
## 1 0.479     1     0    0.292  0.196     0     0.5   0.740     0    0.161
## 2 0.5       1     0    0.434  0.176     1     0     0.641     1    0.5
## 3 0.854     1     0    0.481  0.110     0     0.5   0.412     1    0.419
## 4 0.667     1     0    0.509  0.176     0     0.5   0.687     0    0
## 5 0.688     0     0    0.415  0.384     1     0.5   0.267     0    0.306
## 6 0.604     0     0    0.0566 0.279     0     0     0.389     0    0.161
## # ... with 4 more variables: slope <dbl>, ca <dbl>, thal <dbl>,
## #   target <dbl>
```

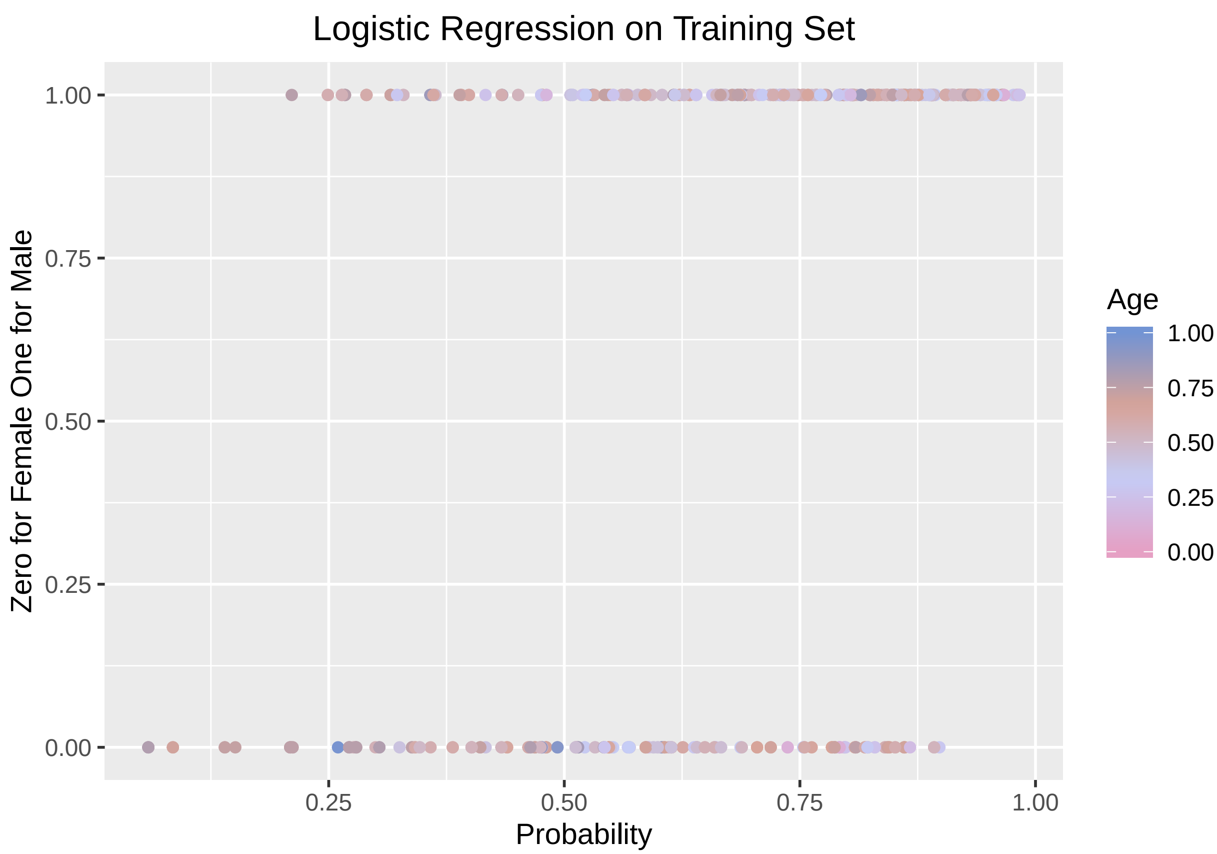
Training Error

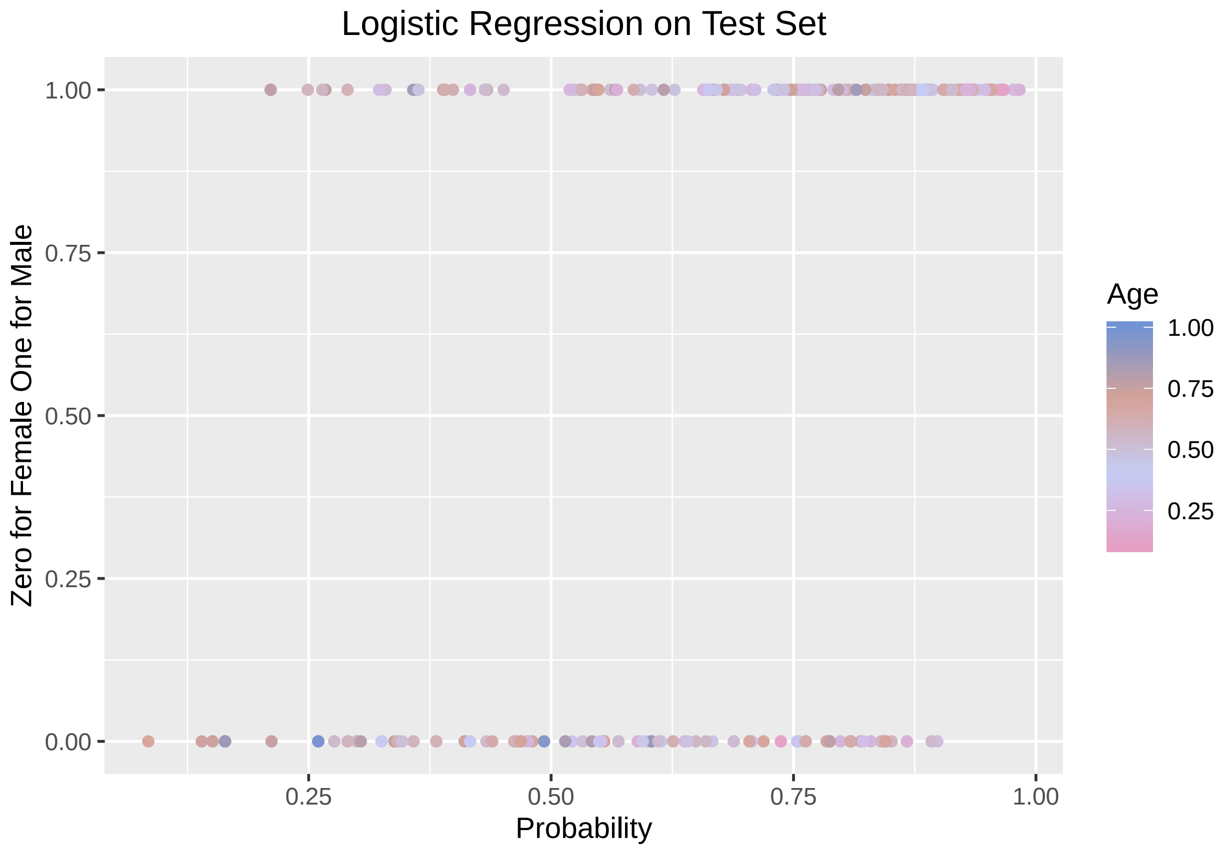
```
##
## predict60   0   1
##         0  68  46
##         1 139 445
```

```
## [1] 0.265043
```

Testing Error

```
## [1] 0.2691131
```





Overall Conclusion

Logistic regression perform the best when it came to classifying gender based on the other variables in the data set. The neural network I believe would outperform the logistic regression if more nodes and hidden layers are added because of the further tuning of weights that would happen if they were added. PCA helped me visually understand what going on between the variables in the data set. Such predictors like cp, threstbp, and chol are good predictors

Roadblocks and Struggles

Learning PCA in a more in-depth was a challenge and still not satisfied with my understanding of it. The neural network was the same. This was a completely new concept to myself. There were a lot of concepts in both models that I would have to grasp. This was a fun project to work on and I will do updates to this analysis in the future.

References

https://www.youtube.com/watch?v=Ilg3gGewQ5U&list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi&index=3

https://www.datanovia.com/en/blog/top-r-color-palettes-to-know-for-great-data-visualization/

https://datascience.stackexchange.com/questions/13178/how-to-normalize-data-for-neural-network-and-decision-forest

https://www.youtube.com/watch?v=-Vs9Vae2KI0&list=LLzI53HRURuRX0Whnv35jvcA&index=12&t=0s

https://www.youtube.com/watch?v=aircAruvnKk&list=LLzI53HRURuRX0Whnv35jvcA&index=11&t=0s

https://www.youtube.com/watch?v=OowGKNgdowA&list=LLzI53HRURuRX0Whnv35jvcA&index=21&t=0s

Introduction to Statistical Learning

Elements of Statistical learning

Data Mining Ian H Witten