Exercise 13

Taniya Adhikari"

10/25/2020

## **Data Set**

Table 13.1: A table of the Dataset with first 5 values.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DGN | PRE4 | PRE5 | PRE6 | PRE7 | PRE8 | PRE9 | PRE10 | PRE11 | PRE14 | PRE17 | PRE19 | PRE25 | PRE30 | PRE32 | AGE | Risk1Yr |
| DGN2 | 2.88 | 2.16 | PRZ1 | F | F | F | T | T | OC14 | F | F | F | T | F | 60 | F |
| DGN3 | 3.40 | 1.88 | PRZ0 | F | F | F | F | F | OC12 | F | F | F | T | F | 51 | F |
| DGN3 | 2.76 | 2.08 | PRZ1 | F | F | F | T | F | OC11 | F | F | F | T | F | 59 | F |
| DGN3 | 3.68 | 3.04 | PRZ0 | F | F | F | F | F | OC11 | F | F | F | F | F | 54 | F |
| DGN3 | 2.44 | 0.96 | PRZ2 | F | T | F | T | T | OC11 | F | F | F | T | F | 73 | T |

## **Summary Statistics**

##   
## Call:  
## glm(formula = Risk1Yr ~ ., family = binomial, data = surgery\_df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6084 -0.5439 -0.4199 -0.2762 2.4929   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.655e+01 2.400e+03 -0.007 0.99450   
## DGNDGN2 1.474e+01 2.400e+03 0.006 0.99510   
## DGNDGN3 1.418e+01 2.400e+03 0.006 0.99528   
## DGNDGN4 1.461e+01 2.400e+03 0.006 0.99514   
## DGNDGN5 1.638e+01 2.400e+03 0.007 0.99455   
## DGNDGN6 4.089e-01 2.673e+03 0.000 0.99988   
## DGNDGN8 1.803e+01 2.400e+03 0.008 0.99400   
## PRE4 -2.272e-01 1.849e-01 -1.229 0.21909   
## PRE5 -3.030e-02 1.786e-02 -1.697 0.08971 .   
## PRE6PRZ1 -4.427e-01 5.199e-01 -0.852 0.39448   
## PRE6PRZ2 -2.937e-01 7.907e-01 -0.371 0.71030   
## PRE7T 7.153e-01 5.556e-01 1.288 0.19788   
## PRE8T 1.743e-01 3.892e-01 0.448 0.65419   
## PRE9T 1.368e+00 4.868e-01 2.811 0.00494 \*\*  
## PRE10T 5.770e-01 4.826e-01 1.196 0.23185   
## PRE11T 5.162e-01 3.965e-01 1.302 0.19295   
## PRE14OC12 4.394e-01 3.301e-01 1.331 0.18318   
## PRE14OC13 1.179e+00 6.165e-01 1.913 0.05580 .   
## PRE14OC14 1.653e+00 6.094e-01 2.713 0.00668 \*\*  
## PRE17T 9.266e-01 4.445e-01 2.085 0.03709 \*   
## PRE19T -1.466e+01 1.654e+03 -0.009 0.99293   
## PRE25T -9.789e-02 1.003e+00 -0.098 0.92227   
## PRE30T 1.084e+00 4.990e-01 2.172 0.02984 \*   
## PRE32T -1.398e+01 1.645e+03 -0.008 0.99322   
## AGE -9.506e-03 1.810e-02 -0.525 0.59944   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 395.61 on 469 degrees of freedom  
## Residual deviance: 341.19 on 445 degrees of freedom  
## AIC: 391.19  
##   
## Number of Fisher Scoring iterations: 15

According to the summary, variable PRE9 - Dyspnoea before surgery being True and PRE14 - size of the original tumour OC14 are highly significant variables with p-value < 0.01 and has greater impact of Risk1Yr being True. PRE17 - Type 2 DM - diabetes mellitus being True and PRE30 - Smoking being True are also significant with p-value < 0.05. All the other variables has p-value between 0.05 < p-value < 1, indicating lesser significance. The value of the deviance for this model (341.19) is also less than the null model (395.61).

|  |  |  |
| --- | --- | --- |
|  | FALSE | TRUE |
| **F** | 390 | 10 |
| **T** | 67 | 3 |

It seems like the model is good at predicting false values, with the accuracy rate of 85.3%. However the accuracy rate of True values of predicting the 23.07%. Suggests that model needs improvement, maybe needs dimensionality reduction.