HW3 STATS 402

Clayton Chan

2024-12-06

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

11.

```
## Generalized Linear Model
##
## 12453 samples
##     4 predictor
##     2 classes: 'Failure', 'Success'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 9962, 9963, 9962, 9963, 9962
## Resampling results:
##
##   Accuracy   Kappa
##   0.8067134  0.09673407
```

From doing the 5-fold cross validation, our estimated accuracy for the logistic regression model fitted on these 4 variables is estimated to be roughly 80.7% which is pretty decent on the surface but the kappa suggests otherwise since it is pretty low at 9.7%. This could be due to how imbalanced this dataset is which the kappa value handles quite well as opposed to accuracy and thus we conclude that this model performs quite poorly.

12a. The question of interest is whether there is an interaction effect between the ethnicity of the donor and height of the donor?

b. The answer is no as the interaction plots look very parallel indicating a lack of significant interaction effect and we can later confirm this in the output once we fit the model.

13.

```
##
## Call:
## glm(formula = tx_fail ~ . + hgt_cm_don_calc.x * ethnicity_don,
##     family = "binomial", data = df)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                      -2.890760   0.453066  -6.380 1.77e-10 ***
## hgt_cm_don_calc.x                 0.027789   0.002524  11.012  < 2e-16 ***
## bmi_don_calc.x                   -0.010206   0.004398  -2.321   0.0203 *
## coronary_angio_don.xY             0.715012   0.072129   9.913  < 2e-16 ***
## hist_hypertens_don.xY            -1.316722   0.053600 -24.566  < 2e-16 ***
## ethnicity_don                     1.256240   1.191557   1.054   0.2918
## hgt_cm_don_calc.x:ethnicity_don -0.006946   0.007039  -0.987   0.3238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12358  on 12452  degrees of freedom
## Residual deviance: 11450  on 12446  degrees of freedom
## AIC: 11464
##
## Number of Fisher Scoring iterations: 4
```

The output shows that the interaction effect is not significant as the p-value .3238 for the respective coefficient is greater than our significance level of 0.05 and thus we FTR the null hypothesis and this further solidifies our claims in 12.

14.

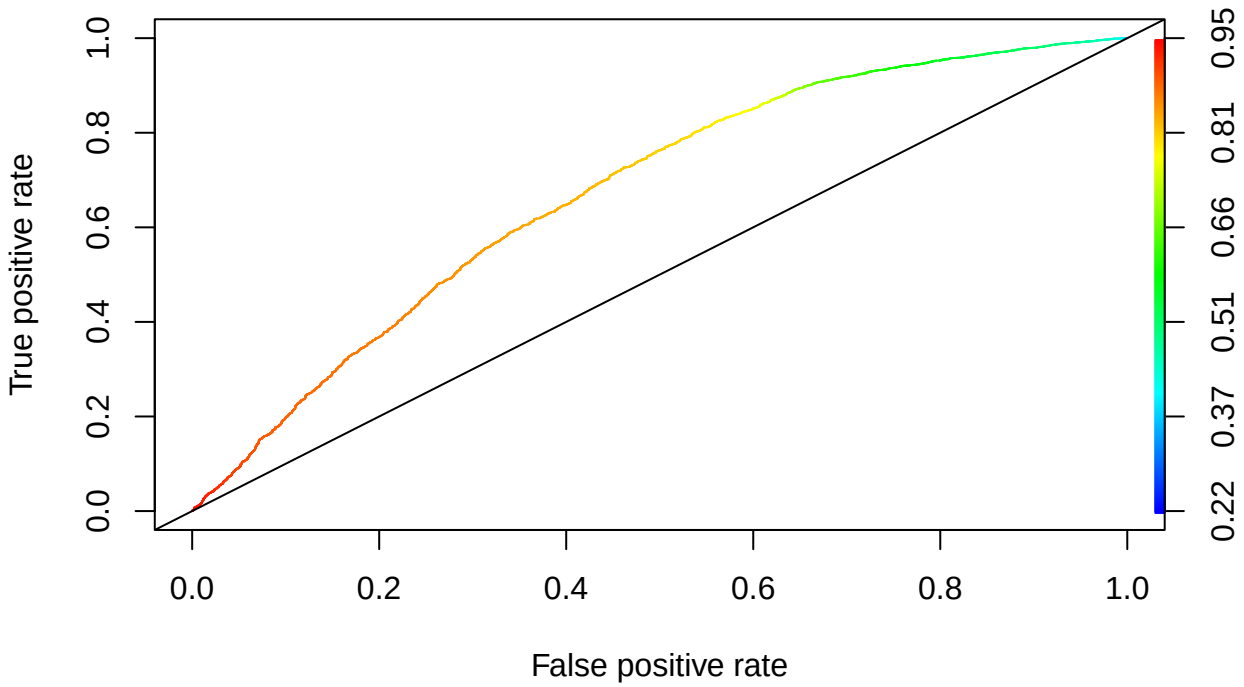
```
##
## Call:
## glm(formula = tx_fail ~ . - ethnicity_don, family = "binomial",
##     data = df)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.671619   0.422976  -6.316 2.68e-10 ***
## hgt_cm_don_calc.x      0.026568   0.002343  11.340  < 2e-16 ***
## bmi_don_calc.x        -0.010079   0.004400  -2.291    0.022 *
## coronary_angio_don.xY  0.714100   0.072138   9.899  < 2e-16 ***
## hist_hypertens_don.xY -1.320632   0.053513 -24.679  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12358  on 12452  degrees of freedom
## Residual deviance: 11452  on 12448  degrees of freedom
## AIC: 11462
##
## Number of Fisher Scoring iterations: 4
```

```
## Analysis of Deviance Table
##
## Model 1: tx_fail ~ (hgt_cm_don_calc.x + bmi_don_calc.x + coronary_angio_don.x +
##     hist_hypertens_don.x + ethnicity_don) - ethnicity_don
## Model 2: tx_fail ~ hgt_cm_don_calc.x + bmi_don_calc.x + coronary_angio_don.x +
##     hist_hypertens_don.x + ethnicity_don + hgt_cm_don_calc.x *
```

```
##     ethnicity_don
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     12448      11452
## 2     12446      11450  2   2.4547   0.2931
```

The null hypothesis for this ANOVA test is that $H_0 : \beta_{eth} = 0, \beta_{hgt*eth} = 0$ with $H_a : \beta_{eth} \neq 0$ or $\beta_{hgt*eth} \neq 0$. Our output shows that the p-value for this test .2931 is greater than our significance level of 0.05 meaning that we FTR the null hypothesis and can drop the ethnicity_don variable alongside the interaction term and as a result, I would recommend the first model model.

15.




```
## [1] 0.6737742
```

The AUC is roughly .67 which means that the model doesn't have that great of a performance. While it's quite far from 1 (which is perfect performance) ,it does slightly better than random guessing.

```
knitr::opts_chunk$set(echo = F)
library(caret)
set.seed(123)
df <- read.csv("liver23.csv")
df <- df[,c("hgt_cm_don_calc.x","bmi_don_calc.x","coronary_angio_don.x","hist_hypertens_don.x","tx_fail
df <- df[df$coronary_angio_don.x!="U" & df$hist_hypertens_don.x!="U",] #Remove unknown level
df$tx_fail <- factor(ifelse(df$tx_fail==1,"Failure","Success")) #Rename outcome and set reference as fa
df$tx_fail <- relevel(df$tx_fail,ref="Failure") #So model predicts probabiltiy for success
df$coronary_angio_don.x <- factor(df$coronary_angio_don.x)
df$hist_hypertens_don.x <- factor(df$hist_hypertens_don.x)
library(caret)
set.seed(123)
ctrlspecs <- trainControl(method="cv", number=5, classProbs=TRUE) #Set up and fit 5 fold cv
```

```r
model1 <- train(tx_fail ~  ., data=df,
                method="glm",
                family=binomial,
                trControl=ctrlspecs)
print(model1)
set.seed(123)
df <- read.csv("liver23.csv")
df <- df[,c("hgt_cm_don_calc.x","bmi_don_calc.x","coronary_angio_don.x","hist_hypertens_don.x","tx_fail
df <- df[df$coronary_angio_don.x!="U" & df$hist_hypertens_don.x!="U",]
df$tx_fail <- factor(ifelse(df$tx_fail==1,"Failure","Success"))
df$tx_fail <- relevel(df$tx_fail,ref="Failure")
df$coronary_angio_don.x <- factor(df$coronary_angio_don.x)
df$hist_hypertens_don.x <- factor(df$hist_hypertens_don.x)

logreg2 <- glm(tx_fail ~ .+hgt_cm_don_calc.x*ethnicity_don,family = "binomial", data = df)
summary(logreg2) #Fit logistic regression with interaction
logreg1 <- glm(tx_fail ~ . -ethnicity_don,family = "binomial", data = df)
summary(logreg1)
anova(logreg1,logreg2,test="Chisq") #Compare model 1 with 4 terms and model 2 with 6 terms including in
library(ROCR)
probpredictions <- predict(logreg1, newdata=df, type="response") #Retrieve probabilities to make ROC cu
pred_m1 <- prediction(probpredictions, df$tx_fail)
roc_curve <- performance(pred_m1, "tpr","fpr")
plot(roc_curve, colorize=T)
abline(0, 1)
auc_ROCR <- performance(pred_m1, measure = "auc")
(auc_ROCR <- auc_ROCR@y.values[[1]]) #Retrieve AUC
```