HW3 STATS 402

Clayton Chan

2024-12-06

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

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

11.

```
## Generalized Linear Model
##
## 9963 samples
##    4 predictor
##    2 classes: 'Success', 'Failure'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 7970, 7971, 7970, 7971, 7970
## Resampling results:
##
##   Accuracy   Kappa
##   0.8071862  0.1045625
```

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 10.4%. 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 = train_df)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                      3.136126   0.509197   6.159 7.32e-10 ***
## hgt_cm_don_calc.x               -0.028697   0.002840 -10.106  < 2e-16 ***
## bmi_don_calc.x                   0.006691   0.004900   1.365    0.172
## coronary_angio_don.xY           -0.659729   0.079429  -8.306  < 2e-16 ***
## hist_hypertens_don.xY            1.336192   0.060049  22.252  < 2e-16 ***
## ethnicity_don                   -1.023971   1.314465  -0.779    0.436
## hgt_cm_don_calc.x:ethnicity_don  0.005307   0.007774   0.683    0.495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9888.4  on 9962  degrees of freedom
## Residual deviance: 9154.5  on 9956  degrees of freedom
## AIC: 9168.5
##
## Number of Fisher Scoring iterations: 4
```

The output shows that the interaction effect is not significant as the p-value .495 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.

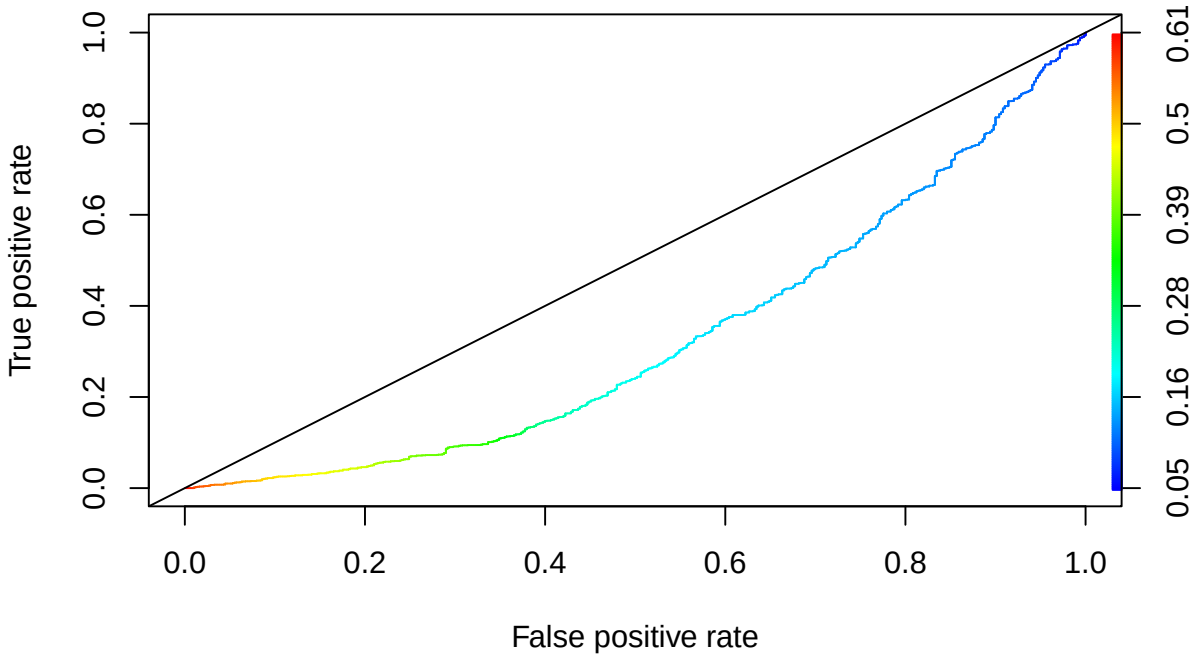
14.

```
##
## Call:
## glm(formula = tx_fail ~ . - ethnicity_don, family = "binomial",
##     data = train_df)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             2.906580   0.473807   6.135 8.54e-10 ***
## hgt_cm_don_calc.x      -0.027446   0.002627 -10.450  < 2e-16 ***
## bmi_don_calc.x          0.006498   0.004904   1.325    0.185
## coronary_angio_don.xY -0.658675   0.079453  -8.290  < 2e-16 ***
## hist_hypertens_don.xY  1.342476   0.059952  22.393  < 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: 9888.4  on 9962  degrees of freedom
## Residual deviance: 9157.9  on 9958  degrees of freedom
## AIC: 9167.9
##
## 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      9958     9157.9
## 2      9956     9154.5  2   3.3139   0.1907
```

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 .1907 is greater than our significance level of 0.05 meaning that we FTR the null hypothesis and can drop the ethnicity variable alongside the interaction term and as a result, I would recommend the first model model.



15.

```
## [1] 0.3300469
```

The AUC is roughly .33 which means that the model performs very poorly as it's very far away from 1 which is perfect performance.

```
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$hist_hypertens_don.x!="U",]
df$tx_fail <- factor(ifelse(df$tx_fail==1,"Failure","Success"))
df$tx_fail <- relevel(df$tx_fail,ref="Success")
df$coronary_angio_don.x <- factor(df$coronary_angio_don.x)
df$hist_hypertens_don.x <- factor(df$hist_hypertens_don.x)
index <- createDataPartition(df$tx_fail, p=.8, list=FALSE, times=1)
train_df <- df[index,]
test_df <- df[-index,]
library(caret)
set.seed(123)
ctrlspecs <- trainControl(method="cv", number=5, classProbs=TRUE)
```

```
model1 <- train(tx_fail ~  ., data=train_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="Success")
df$coronary_angio_don.x <- factor(df$coronary_angio_don.x)
df$hist_hypertens_don.x <- factor(df$hist_hypertens_don.x)
index <- createDataPartition(df$tx_fail, p=.8, list=FALSE, times=1)
train_df <- df[index,]
test_df <- df[-index,]
logreg2 <- glm(tx_fail ~ .+hgt_cm_don_calc.x*ethnicity_don,family = "binomial", data = train_df)
summary(logreg2)
logreg1 <- glm(tx_fail ~ . -ethnicity_don,family = "binomial", data = train_df)
summary(logreg1)
anova(logreg1,logreg2,test="Chisq")
library(ROCR)
ctrlspecs <- trainControl(method="cv", number=5, classProbs=TRUE)
model2 <- train(tx_fail ~  . - ethnicity_don, data=train_df,
                method="glm",
                family=binomial,
                trControl=ctrlspecs)
probpredictions <- predict(model2, newdata=test_df, type="prob")
probpredictions <- probpredictions[,"Failure"]
pred_m1 <- prediction(probpredictions, test_df$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]])
```