Homework5

Justin Markey

10/26/2021

# Import the data

defaultdata <- read.csv("data/HW5 Data.csv")  
defaultdata$SEX <- factor(defaultdata$SEX)  
defaultdata$EDUCATION <- factor(defaultdata$EDUCATION)  
defaultdata$MARRIAGE <- factor(defaultdata$MARRIAGE)  
summary(defaultdata)

## LIMIT\_BAL SEX EDUCATION MARRIAGE AGE   
## Min. : 10000 1:11888 0: 14 0: 54 Min. :21.00   
## 1st Qu.: 50000 2:18112 1:10585 1:13659 1st Qu.:28.00   
## Median : 140000 2:14030 2:15964 Median :34.00   
## Mean : 167484 3: 4917 3: 323 Mean :35.49   
## 3rd Qu.: 240000 4: 123 3rd Qu.:41.00   
## Max. :1000000 5: 280 Max. :79.00   
## 6: 51   
## BILL.AMT PAY.AMT DEFAULT   
## Min. :-165580 Min. : 0 Min. :0.0000   
## 1st Qu.: 3559 1st Qu.: 1000 1st Qu.:0.0000   
## Median : 22382 Median : 2100 Median :0.0000   
## Mean : 51223 Mean : 5664 Mean :0.2212   
## 3rd Qu.: 67091 3rd Qu.: 5006 3rd Qu.:0.0000   
## Max. : 964511 Max. :873552 Max. :1.0000   
##

# Q1

## Data partiiton

set.seed(12345)  
row\_ind<-sample(1:nrow(defaultdata),20000)  
traindata<-defaultdata[row\_ind,]  
validdata<-defaultdata[-row\_ind,]  
"Train"

## [1] "Train"

summary(traindata)

## LIMIT\_BAL SEX EDUCATION MARRIAGE AGE   
## Min. : 10000 1: 7954 0: 13 0: 37 Min. :21.00   
## 1st Qu.: 50000 2:12046 1:7068 1: 9099 1st Qu.:28.00   
## Median :140000 2:9346 2:10658 Median :34.00   
## Mean :168078 3:3277 3: 206 Mean :35.51   
## 3rd Qu.:240000 4: 75 3rd Qu.:41.00   
## Max. :780000 5: 184 Max. :79.00   
## 6: 37   
## BILL.AMT PAY.AMT DEFAULT   
## Min. :-154973 Min. : 0 Min. :0.0000   
## 1st Qu.: 3537 1st Qu.: 1000 1st Qu.:0.0000   
## Median : 22176 Median : 2100 Median :0.0000   
## Mean : 50993 Mean : 5573 Mean :0.2192   
## 3rd Qu.: 66096 3rd Qu.: 5003 3rd Qu.:0.0000   
## Max. : 746814 Max. :873552 Max. :1.0000   
##

"Valid"

## [1] "Valid"

summary(validdata)

## LIMIT\_BAL SEX EDUCATION MARRIAGE AGE   
## Min. : 10000 1:3934 0: 1 0: 17 Min. :21.00   
## 1st Qu.: 50000 2:6066 1:3517 1:4560 1st Qu.:28.00   
## Median : 140000 2:4684 2:5306 Median :34.00   
## Mean : 166296 3:1640 3: 117 Mean :35.44   
## 3rd Qu.: 240000 4: 48 3rd Qu.:41.00   
## Max. :1000000 5: 96 Max. :75.00   
## 6: 14   
## BILL.AMT PAY.AMT DEFAULT   
## Min. :-165580 Min. : 0.0 Min. :0.0000   
## 1st Qu.: 3602 1st Qu.: 992.5 1st Qu.:0.0000   
## Median : 22936 Median : 2128.0 Median :0.0000   
## Mean : 51684 Mean : 5844.8 Mean :0.2251   
## 3rd Qu.: 69362 3rd Qu.: 5039.2 3rd Qu.:0.0000   
## Max. : 964511 Max. :505000.0 Max. :1.0000   
##

# Q2

## logit model without age

defaultmodel <- glm(data = traindata, DEFAULT ~ LIMIT\_BAL+SEX+EDUCATION+MARRIAGE+BILL.AMT+PAY.AMT, family = binomial)  
summary(defaultmodel)

##   
## Call:  
## glm(formula = DEFAULT ~ LIMIT\_BAL + SEX + EDUCATION + MARRIAGE +   
## BILL.AMT + PAY.AMT, family = binomial, data = traindata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0145 -0.7684 -0.6360 -0.3566 4.3812   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.325e+01 8.832e+01 -0.150 0.8807   
## LIMIT\_BAL -3.375e-06 1.777e-07 -18.989 < 2e-16 \*\*\*  
## SEX2 -1.987e-01 3.536e-02 -5.620 1.91e-08 \*\*\*  
## EDUCATION1 1.102e+01 8.832e+01 0.125 0.9007   
## EDUCATION2 1.107e+01 8.832e+01 0.125 0.9003   
## EDUCATION3 1.101e+01 8.832e+01 0.125 0.9008   
## EDUCATION4 9.532e+00 8.832e+01 0.108 0.9141   
## EDUCATION5 9.592e+00 8.832e+01 0.109 0.9135   
## EDUCATION6 1.075e+01 8.832e+01 0.122 0.9031   
## MARRIAGE1 1.767e+00 7.311e-01 2.417 0.0156 \*   
## MARRIAGE2 1.470e+00 7.313e-01 2.009 0.0445 \*   
## MARRIAGE3 1.837e+00 7.467e-01 2.460 0.0139 \*   
## BILL.AMT 1.798e-06 2.857e-07 6.293 3.12e-10 \*\*\*  
## PAY.AMT -2.659e-05 3.271e-06 -8.130 4.29e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21038 on 19999 degrees of freedom  
## Residual deviance: 20199 on 19986 degrees of freedom  
## AIC: 20227  
##   
## Number of Fisher Scoring iterations: 11

### AIC is

## [1] 20226.68

# Q3

1. increase payment: with an increase in PAY.AMT, You are able to REDUCE defaulting by

## PAY.AMT   
## -2.659258e-05

1. increase bill: with an increase in the bill, you INCREASE defaulting by

## BILL.AMT   
## 1.797871e-06

1. increase limit balance: with an increase in limit balance, you are able to REDUCE defaulting by

## LIMIT\_BAL   
## -3.37445e-06

1. A decrease in marriage status by 2 will INCREASE defaulting by

## MARRIAGE3   
## 0.06938744

# Q4

## McFadden Pseudo R-Squared

LL.full<-as.numeric(logLik(defaultmodel))  
null.defaultmodel<-glm(data=defaultdata,  
 family=binomial,  
 DEFAULT ~ 1)  
LL.null<-as.numeric(logLik(null.defaultmodel))  
  
mcfaddenrsqr <- 1-(LL.full)/(LL.null)

### Mcfadden Pseudo R-Square is equal to

## [1] 0.3629254

# Q5

## set data in dataframe

predictiondata <- data.frame(LIMIT\_BAL = c(50000),  
 SEX = c(as.factor(1)),  
 EDUCATION = c(as.factor(5)),  
 MARRIAGE = c(as.factor(3)),  
 BILL.AMT = c(2500),  
 PAY.AMT = c(1000)  
 )

### Conclusion

The percentage of default is

predict(defaultmodel, newdata = predictiondata, type = "response")

## 1   
## 0.1174417

# Q6

## make model with age in it

ageindefaultmodel<- glm(data=traindata, DEFAULT ~ ., family=binomial)  
summary(ageindefaultmodel)

##   
## Call:  
## glm(formula = DEFAULT ~ ., family = binomial, data = traindata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0194 -0.7673 -0.6359 -0.3573 4.3797   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.338e+01 8.837e+01 -0.151 0.8796   
## LIMIT\_BAL -3.405e-06 1.789e-07 -19.030 < 2e-16 \*\*\*  
## SEX2 -1.920e-01 3.564e-02 -5.389 7.09e-08 \*\*\*  
## EDUCATION1 1.103e+01 8.837e+01 0.125 0.9006   
## EDUCATION2 1.108e+01 8.837e+01 0.125 0.9002   
## EDUCATION3 1.101e+01 8.837e+01 0.125 0.9008   
## EDUCATION4 9.553e+00 8.837e+01 0.108 0.9139   
## EDUCATION5 9.608e+00 8.837e+01 0.109 0.9134   
## EDUCATION6 1.074e+01 8.837e+01 0.122 0.9033   
## MARRIAGE1 1.755e+00 7.311e-01 2.400 0.0164 \*   
## MARRIAGE2 1.483e+00 7.313e-01 2.028 0.0426 \*   
## MARRIAGE3 1.815e+00 7.468e-01 2.430 0.0151 \*   
## AGE 3.244e-03 2.149e-03 1.510 0.1311   
## BILL.AMT 1.792e-06 2.856e-07 6.274 3.51e-10 \*\*\*  
## PAY.AMT -2.655e-05 3.270e-06 -8.121 4.64e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 21038 on 19999 degrees of freedom  
## Residual deviance: 20196 on 19985 degrees of freedom  
## AIC: 20226  
##   
## Number of Fisher Scoring iterations: 11

## Compare AICs

"model with age"

## [1] "model with age"

AIC(ageindefaultmodel)

## [1] 20226.4

"model without age"

## [1] "model without age"

AIC(defaultmodel)

## [1] 20226.68

### Conclusion

The model with age included yielded a lower AIC. The lower the AIC the better as it has a built in cost mechanism, similar to adj R Square in linear regressions. Adding more variables to a model adds robustness, generally speaking. The inclusion of age was not only helpful to the model performance, but allows the model to take in more data in general, and perhaps be a better predictor of defaulting in this case.