Report

2022-11-21

library(readr)  
# Load in health Data  
health <- read\_csv("C:/Users/Alyssa Guillory/Desktop/MATH 4322/GROUP PROJ/Project/ML-Group7/healthcare-dataset-stroke-data.csv", col\_types = cols(gender = col\_factor(levels = c("Male", "Female", "Other")), hypertension = col\_factor(levels = c("0", "1")), heart\_disease = col\_factor(levels = c("0", "1")), ever\_married = col\_factor(levels = c("No", "Yes")), work\_type = col\_factor(levels = c("children", "Govt\_job", "Never\_worked", "Private", "Self-employed")), Residence\_type = col\_factor(levels = c("Rural", "Urban")), bmi = col\_number(), smoking\_status = col\_factor(levels = c("formerly smoked", "never smoked", "smokes")), stroke = col\_factor(levels = c("0", "1"))))  
# Remove ID column  
health = health[,-1]  
# Omit NA values  
health2 = na.omit(health)

#create model with all predictors  
set.seed(500)  
fit.health = glm(stroke ~ ., family = "binomial", data=health2)  
summary(fit.health)

##   
## Call:  
## glm(formula = stroke ~ ., family = "binomial", data = health2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.2163 -0.3357 -0.1914 -0.1051 3.1357   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.911e+01 4.775e+02 -0.040 0.968076   
## genderFemale 7.625e-02 1.681e-01 0.454 0.650091   
## genderOther -1.234e+01 3.956e+03 -0.003 0.997511   
## age 7.304e-02 6.958e-03 10.497 < 2e-16 \*\*\*  
## hypertension1 5.694e-01 1.828e-01 3.114 0.001845 \*\*   
## heart\_disease1 3.906e-01 2.211e-01 1.767 0.077249 .   
## ever\_marriedYes -1.795e-01 2.627e-01 -0.683 0.494366   
## work\_typeGovt\_job 1.092e+01 4.775e+02 0.023 0.981747   
## work\_typeNever\_worked -2.958e-01 1.157e+03 0.000 0.999796   
## work\_typePrivate 1.117e+01 4.775e+02 0.023 0.981339   
## work\_typeSelf-employed 1.081e+01 4.775e+02 0.023 0.981936   
## Residence\_typeUrban -4.830e-03 1.624e-01 -0.030 0.976276   
## avg\_glucose\_level 4.696e-03 1.377e-03 3.409 0.000651 \*\*\*  
## bmi 6.501e-03 1.292e-02 0.503 0.614789   
## smoking\_statusnever smoked -7.476e-02 1.894e-01 -0.395 0.692999   
## smoking\_statussmokes 3.151e-01 2.305e-01 1.367 0.171639   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1411.0 on 3425 degrees of freedom  
## Residual deviance: 1141.8 on 3410 degrees of freedom  
## AIC: 1173.8  
##   
## Number of Fisher Scoring iterations: 16

#create model with significant predictors found from doing stepwise backward method on original model  
step(fit.health, direction = "backward")

## Start: AIC=1173.81  
## stroke ~ gender + age + hypertension + heart\_disease + ever\_married +   
## work\_type + Residence\_type + avg\_glucose\_level + bmi + smoking\_status  
##   
## Df Deviance AIC  
## - work\_type 4 1145.9 1169.9  
## - gender 2 1142.0 1170.0  
## - Residence\_type 1 1141.8 1171.8  
## - bmi 1 1142.1 1172.1  
## - ever\_married 1 1142.3 1172.3  
## - smoking\_status 2 1145.0 1173.0  
## <none> 1141.8 1173.8  
## - heart\_disease 1 1144.8 1174.8  
## - hypertension 1 1151.0 1181.0  
## - avg\_glucose\_level 1 1153.1 1183.1  
## - age 1 1277.2 1307.2  
##   
## Step: AIC=1169.88  
## stroke ~ gender + age + hypertension + heart\_disease + ever\_married +   
## Residence\_type + avg\_glucose\_level + bmi + smoking\_status  
##   
## Df Deviance AIC  
## - gender 2 1146.1 1166.1  
## - Residence\_type 1 1145.9 1167.9  
## - bmi 1 1146.2 1168.2  
## - ever\_married 1 1146.2 1168.2  
## - smoking\_status 2 1149.2 1169.2  
## <none> 1145.9 1169.9  
## - heart\_disease 1 1149.3 1171.3  
## - hypertension 1 1155.0 1177.0  
## - avg\_glucose\_level 1 1157.2 1179.2  
## - age 1 1283.2 1305.2  
##   
## Step: AIC=1166.06  
## stroke ~ age + hypertension + heart\_disease + ever\_married +   
## Residence\_type + avg\_glucose\_level + bmi + smoking\_status  
##   
## Df Deviance AIC  
## - Residence\_type 1 1146.1 1164.1  
## - bmi 1 1146.4 1164.4  
## - ever\_married 1 1146.5 1164.5  
## - smoking\_status 2 1149.3 1165.3  
## <none> 1146.1 1166.1  
## - heart\_disease 1 1149.4 1167.4  
## - hypertension 1 1155.2 1173.2  
## - avg\_glucose\_level 1 1157.2 1175.2  
## - age 1 1283.8 1301.8  
##   
## Step: AIC=1164.06  
## stroke ~ age + hypertension + heart\_disease + ever\_married +   
## avg\_glucose\_level + bmi + smoking\_status  
##   
## Df Deviance AIC  
## - bmi 1 1146.4 1162.4  
## - ever\_married 1 1146.5 1162.5  
## - smoking\_status 2 1149.3 1163.3  
## <none> 1146.1 1164.1  
## - heart\_disease 1 1149.4 1165.4  
## - hypertension 1 1155.2 1171.2  
## - avg\_glucose\_level 1 1157.2 1173.2  
## - age 1 1283.8 1299.8  
##   
## Step: AIC=1162.36  
## stroke ~ age + hypertension + heart\_disease + ever\_married +   
## avg\_glucose\_level + smoking\_status  
##   
## Df Deviance AIC  
## - ever\_married 1 1146.7 1160.7  
## - smoking\_status 2 1149.6 1161.6  
## <none> 1146.4 1162.4  
## - heart\_disease 1 1149.6 1163.6  
## - hypertension 1 1155.9 1169.9  
## - avg\_glucose\_level 1 1158.9 1172.9  
## - age 1 1286.0 1300.0  
##   
## Step: AIC=1160.74  
## stroke ~ age + hypertension + heart\_disease + avg\_glucose\_level +   
## smoking\_status  
##   
## Df Deviance AIC  
## - smoking\_status 2 1149.9 1159.9  
## <none> 1146.7 1160.7  
## - heart\_disease 1 1150.1 1162.1  
## - hypertension 1 1156.3 1168.3  
## - avg\_glucose\_level 1 1159.1 1171.1  
## - age 1 1293.7 1305.7  
##   
## Step: AIC=1159.91  
## stroke ~ age + hypertension + heart\_disease + avg\_glucose\_level  
##   
## Df Deviance AIC  
## <none> 1149.9 1159.9  
## - heart\_disease 1 1154.1 1162.1  
## - hypertension 1 1159.2 1167.2  
## - avg\_glucose\_level 1 1161.9 1169.9  
## - age 1 1296.1 1304.1

##   
## Call: glm(formula = stroke ~ age + hypertension + heart\_disease + avg\_glucose\_level,   
## family = "binomial", data = health2)  
##   
## Coefficients:  
## (Intercept) age hypertension1 heart\_disease1   
## -7.632810 0.067773 0.568379 0.453704   
## avg\_glucose\_level   
## 0.004701   
##   
## Degrees of Freedom: 3425 Total (i.e. Null); 3421 Residual  
## Null Deviance: 1411   
## Residual Deviance: 1150 AIC: 1160

better.health = glm(formula = stroke ~ age + hypertension + heart\_disease + avg\_glucose\_level,   
 family = "binomial", data = health2)  
  
#create a model w/out heart\_disease  
best.health = glm(formula = stroke ~ age + hypertension + avg\_glucose\_level, family = "binomial", data = health2)

#comparing models  
#reminder: table w/null dev, residual dev, r^2, AIC for each model  
  
#fit.health info  
summary(fit.health)

##   
## Call:  
## glm(formula = stroke ~ ., family = "binomial", data = health2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.2163 -0.3357 -0.1914 -0.1051 3.1357   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.911e+01 4.775e+02 -0.040 0.968076   
## genderFemale 7.625e-02 1.681e-01 0.454 0.650091   
## genderOther -1.234e+01 3.956e+03 -0.003 0.997511   
## age 7.304e-02 6.958e-03 10.497 < 2e-16 \*\*\*  
## hypertension1 5.694e-01 1.828e-01 3.114 0.001845 \*\*   
## heart\_disease1 3.906e-01 2.211e-01 1.767 0.077249 .   
## ever\_marriedYes -1.795e-01 2.627e-01 -0.683 0.494366   
## work\_typeGovt\_job 1.092e+01 4.775e+02 0.023 0.981747   
## work\_typeNever\_worked -2.958e-01 1.157e+03 0.000 0.999796   
## work\_typePrivate 1.117e+01 4.775e+02 0.023 0.981339   
## work\_typeSelf-employed 1.081e+01 4.775e+02 0.023 0.981936   
## Residence\_typeUrban -4.830e-03 1.624e-01 -0.030 0.976276   
## avg\_glucose\_level 4.696e-03 1.377e-03 3.409 0.000651 \*\*\*  
## bmi 6.501e-03 1.292e-02 0.503 0.614789   
## smoking\_statusnever smoked -7.476e-02 1.894e-01 -0.395 0.692999   
## smoking\_statussmokes 3.151e-01 2.305e-01 1.367 0.171639   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1411.0 on 3425 degrees of freedom  
## Residual deviance: 1141.8 on 3410 degrees of freedom  
## AIC: 1173.8  
##   
## Number of Fisher Scoring iterations: 16

fit.bic=BIC(fit.health)  
fit.stat = cbind(fit.health$null.deviance,fit.health$deviance,(1-(fit.health$deviance/fit.health$null.deviance)),fit.health$aic,fit.bic)  
colnames(fit.stat) = c("Null Deviance","Residual Deviance","R^2","AIC","BIC")  
#print(fit.stat)  
#R^2: 1-(1141.8/1411) =0.1907867  
  
#better.health info  
summary(better.health)

##   
## Call:  
## glm(formula = stroke ~ age + hypertension + heart\_disease + avg\_glucose\_level,   
## family = "binomial", data = health2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.1420 -0.3357 -0.1927 -0.1072 3.1976   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.632810 0.439480 -17.368 < 2e-16 \*\*\*  
## age 0.067773 0.006359 10.659 < 2e-16 \*\*\*  
## hypertension1 0.568379 0.181386 3.134 0.001727 \*\*   
## heart\_disease1 0.453704 0.216660 2.094 0.036253 \*   
## avg\_glucose\_level 0.004701 0.001334 3.524 0.000426 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1411.0 on 3425 degrees of freedom  
## Residual deviance: 1149.9 on 3421 degrees of freedom  
## AIC: 1159.9  
##   
## Number of Fisher Scoring iterations: 7

better.bic=BIC(better.health)  
better.stat = cbind(better.health$null.deviance,better.health$deviance,(1-(better.health$deviance/better.health$null.deviance)),better.health$aic,better.bic)  
colnames(better.stat) = c("Null Deviance","Residual Deviance","R^2","AIC","BIC")  
#print(better.stat)  
#R^2: 1-(1149/1411) =0.1856839  
  
#best.health info  
summary(best.health)

##   
## Call:  
## glm(formula = stroke ~ age + hypertension + avg\_glucose\_level,   
## family = "binomial", data = health2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0523 -0.3398 -0.1939 -0.1058 3.2054   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.750124 0.437181 -17.728 < 2e-16 \*\*\*  
## age 0.070099 0.006262 11.194 < 2e-16 \*\*\*  
## hypertension1 0.579247 0.180838 3.203 0.001359 \*\*   
## avg\_glucose\_level 0.004932 0.001326 3.720 0.000199 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1411.0 on 3425 degrees of freedom  
## Residual deviance: 1154.1 on 3422 degrees of freedom  
## AIC: 1162.1  
##   
## Number of Fisher Scoring iterations: 7

best.bic=BIC(best.health)  
best.stat = cbind(best.health$null.deviance,best.health$deviance,(1-(best.health$deviance/best.health$null.deviance)),best.health$aic, best.bic)  
colnames(fit.stat) = c("Null Deviance","Residual Deviance","R^2","AIC","BIC")  
#print(best.stat)  
#R^2: 1-(1154.1/1411) = 0.1820695  
  
goodness.fit=rbind(fit.stat,better.stat,best.stat)  
rownames(goodness.fit) = c("fit.health","better.health","best.health")  
print(goodness.fit)

## Null Deviance Residual Deviance R^2 AIC BIC  
## fit.health 1411.002 1141.806 0.1907835 1173.806 1272.032  
## better.health 1411.002 1149.914 0.1850374 1159.914 1190.609  
## best.health 1411.002 1154.069 0.1820928 1162.069 1186.625

#some info for explaining why we're picking the predictors from better.health:  
#lower residual dev, higher R^2, and lower AIC for better.health shows that including the heart\_disease predictor helps to create a better model  
  
  
#logistic reg. coefficients for equation:  
  
#Intercept:-7.632810  
#Age: 0.067773  
#hypertension1:0.568379  
#heart\_disease1:0.453704  
#avg\_glucose\_level: 0.004701

##training & testing  
test.error = 0  
train.error=0  
for (i in 1:10) {  
 set.seed(i)  
 #Select 80% of the data  
 sample = sample.int(n = nrow(health2),size = round(.80\*nrow(health2)), replace = FALSE)  
   
 train = health2[sample,]  
 test = health2[-sample,]  
 #create model on training data  
 health.train = glm(formula = stroke ~ age + hypertension + heart\_disease + avg\_glucose\_level, family = "binomial", data = train)  
 #summary(health.train)  
   
 #predicting based on the model fit with the training data  
 prediction.train = predict(health.train, type="response")  
 predict.stroke.train = ifelse(prediction.train < 0.5,"0","1")  
 (conf.mat.train = table(predict.stroke.train,train$stroke))  
   
 rows.train = nrow(conf.mat.train)  
 if(rows.train < 2) {  
 train.error[i] = (conf.mat.train[1,2]/sum(conf.mat.train))  
 } else {  
 train.error[i] = (conf.mat.train[1,2]+conf.mat.train[2,1])/sum(conf.mat.train)  
 }  
   
 #predicting based on the model fit with the testing data  
 predict.health = predict(health.train, type = "response", newdata = test)  
 predict.stroke = ifelse(predict.health< 0.5,"0","1")  
 (conf.test = table(predict.stroke,test$stroke))  
   
 rows = nrow(conf.test)  
 if(rows < 2) {  
 test.error[i] = (conf.test[1,2]/sum(conf.test))  
 } else {  
 test.error[i] = (conf.test[1,2]+conf.test[2,1])/sum(conf.test)  
 }  
}  
#Train error rate for each sample:  
train.error

## [1] 0.05290040 0.05326523 0.05435972 0.05107625 0.05034659 0.05144108  
## [7] 0.05144108 0.05472455 0.05435972 0.05363006

#mean of the train error rate  
train.error.mean = mean(train.error)  
train.error.mean

## [1] 0.05275447

#Test error rate for each sample:  
test.error

## [1] 0.05109489 0.04817518 0.04379562 0.05839416 0.06131387 0.05401460  
## [7] 0.05693431 0.04233577 0.04525547 0.04671533

#mean of the test error  
test.error.mean = mean(test.error)  
test.error.mean

## [1] 0.05080292

If the linear regression is used to predict a binary response, the resulting model may not restrict the predicted Y values within 0 and 1.

describe logistic reg choice We chose to use a logistic regression model for our data because our response variable is qualitative with two classes. It was desirable to have a model predict the probability of a person having a stroke using binary classification. If linear regression is used to predict the probability of whether or not a person will have a stroke, then the model may have predicted Y values outside of our intended range of 0-1. The logistic regression model allows us to enforce this restriction of range.

restrict the predicted y values between the set range of 0 and 1.

explain model choice write model equation test/train models test error rate interpretation of full model in regards to our question: which factors demonstrate statistical significance in relation to having a stroke?