Epi\_ML\_Assignment5

Arielle Coq AC4140

2/24/2020

The first thing that I will do is load the data into the R and set the seed to 100. I will create a training and testing data sets that will be used for this analysis. I will also store teh outcome in a separate dataset apart from the other variables.

set.seed(100)  
Alcohol<-read.csv("alcohol\_use.csv")  
  
Alcohol2 = Alcohol %>%   
mutate(alc\_consumption = recode(alc\_consumption, "CurrentUse"= 1, "NotCurrentUse" = 0)) %>%   
 select(-X)  
  
training.data<-Alcohol2$alc\_consumption %>% createDataPartition(p=0.7, list=F)  
train.data1<-Alcohol2[training.data, ]  
test.data1<-Alcohol2[-training.data, ]  
  
#Store outcome   
alc.consumption.train<-train.data1$alc\_consumption  
alc.consumption.test<-test.data1$alc\_consumption  
  
# Store the outcome in one train and test and the predictors in another   
#model matrix- will create indicator variables for categorical varaibles, it does not do anything to the continuous variables  
x.train1<-model.matrix(alc\_consumption~., train.data1)[,-1]  
x.test1<-model.matrix(alc\_consumption~., test.data1)[,-1]

#### Question 1

I will create three different models to determine which model I will use. The three models that I will perform are: a model that chooses alpha and lambda via cross-validation using all of the features, a model that uses all the features and traditional logistic regression, and a lasso model using all of the features

The first model I will do is a model that chooses alpha and lambda via cross-validation using all of the features.

set.seed(100)  
model.1<- train(  
 alc\_consumption ~., data = train.data1, method = "glmnet",  
 trControl = trainControl("cv", number = 10), tuneLength=10, family = "binomial")  
## Tune lenghth = try 10 different parameters -just 10 total   
model.1$bestTune

## alpha lambda  
## 64 0.8 0.266057

model.1a <-glmnet(x.train1, alc.consumption.train, alpha = model.1$bestTune$alpha, lambda = model.1$bestTune$lambda, standardize = TRUE, family = "binomial")

Prediction for the first model

model.1\_fitted<-predict(model.1a,x.test1, type= "response")  
  
fitted.results.p <- ifelse(model.1\_fitted > 0.5,1,0)  
  
testing.model<-(as.numeric(test.data1$alc\_consumption))  
  
model.1\_Error <- mean(fitted.results.p != testing.model, na.rm=T)  
  
  
print(paste('Accuracy Model 1',1-model.1\_Error))

## [1] "Accuracy Model 1 0.844247787610619"

The second model I will do is a model that uses all the features and traditional logistic regression

model.6<-glm(alc\_consumption~., family = binomial(link = "logit"), data = train.data1)  
summary (model.6)

##   
## Call:  
## glm(formula = alc\_consumption ~ ., family = binomial(link = "logit"),   
## data = train.data1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.2067 -0.6528 0.1948 0.7532 1.8684   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.20822 0.07298 2.853 0.00433 \*\*   
## neurotocism\_score 0.09766 0.08639 1.130 0.25829   
## extroversion\_score 0.37749 0.09313 4.054 5.05e-05 \*\*\*  
## openness\_score 0.06913 0.08456 0.818 0.41361   
## agreeableness\_score 0.11549 0.08167 1.414 0.15736   
## conscientiousness\_score -0.07664 0.08546 -0.897 0.36984   
## impulsiveness\_score 1.92183 0.13071 14.703 < 2e-16 \*\*\*  
## sens\_seeking\_score 0.22202 0.10329 2.149 0.03160 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1826.2 on 1319 degrees of freedom  
## Residual deviance: 1171.8 on 1312 degrees of freedom  
## AIC: 1187.8  
##   
## Number of Fisher Scoring iterations: 5

model.6a <-glmnet(x.train1, alc.consumption.train, method = "glm", standardize = TRUE, family = "binomial")

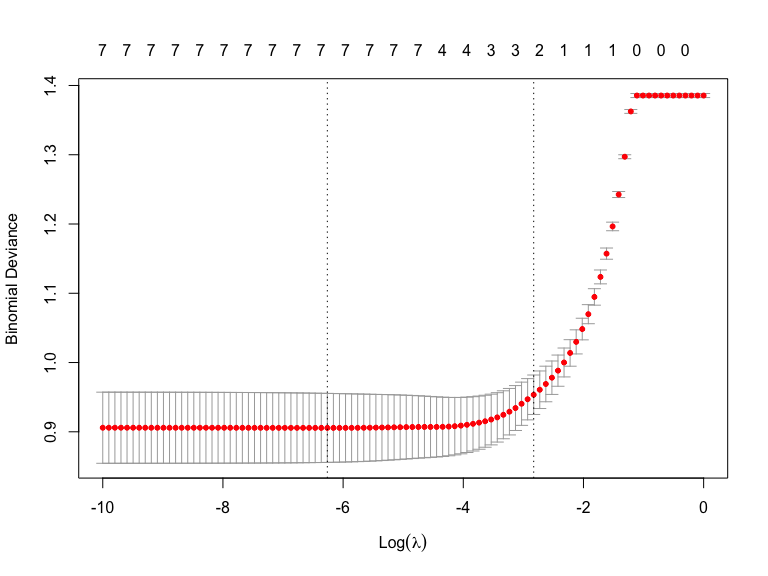
Prediction for model 2

model.6\_fitted<-predict(model.6a, x.test1, type= "response")  
  
fitted.results.p2 <- ifelse(model.6\_fitted > 0.5,1,0)  
  
testing.model.1<-(as.numeric(test.data1$alc\_consumption))  
  
model.6\_Error <- mean(fitted.results.p2 != testing.model.1, na.rm=T)  
  
  
print(paste('Accuracy Model 2',1-model.1\_Error))

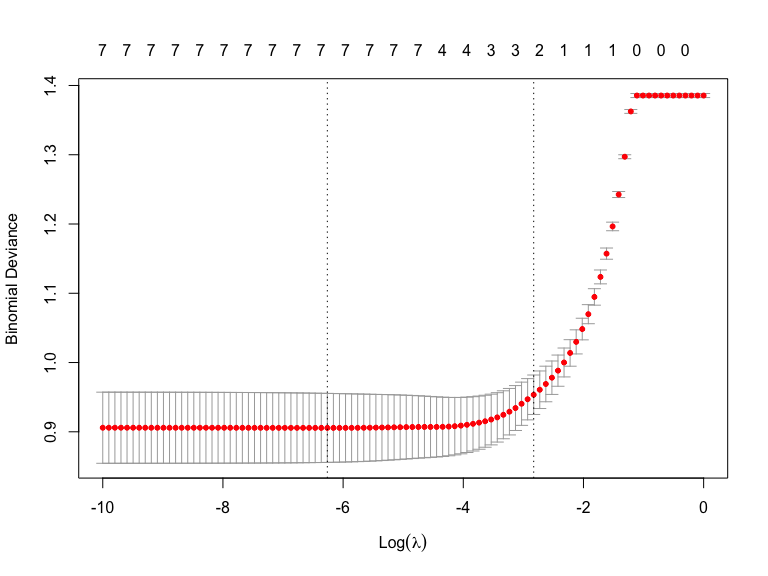
## [1] "Accuracy Model 2 0.844247787610619"

The third and final model that I will do is a lasso model using all of the features

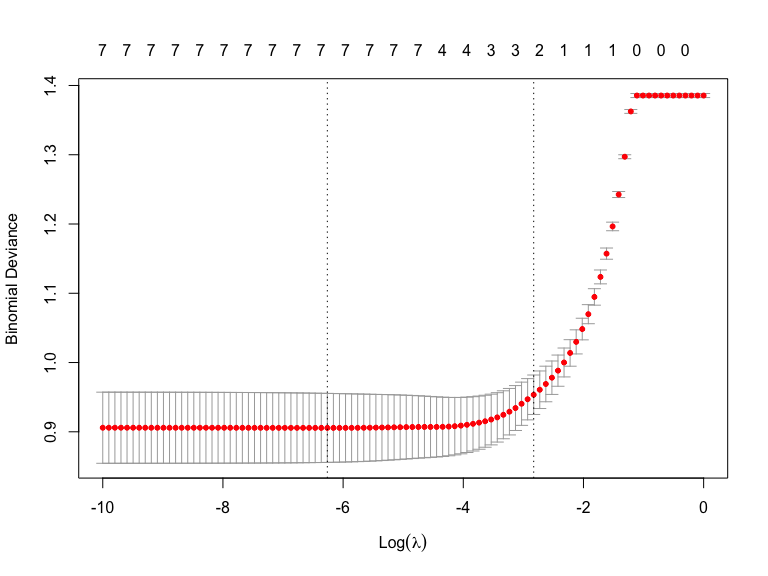
model.7.cv<-cv.glmnet(x.train1, alc.consumption.train, alpha = 1, lambda = exp(seq(-10, 0, length = 100)), family = "binomial")  
  
plot(model.7.cv)



plot(model.7.cv, xvar="lambda", label=TRUE)



plot(model.7.cv, xvar="dev", label=TRUE)



model.7.cv$lambda.min

## [1] 0.001906233

model.7.cv$lambda.1se

## [1] 0.05911428

model.7a<-glmnet(x.train1, alc.consumption.train, alpha=1, family = "binomial", lambda=model.7.cv$lambda.min)  
coef(model.7a)

## 8 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) 0.20512422  
## neurotocism\_score 0.07654444  
## extroversion\_score 0.34984347  
## openness\_score 0.06429381  
## agreeableness\_score 0.09354848  
## conscientiousness\_score -0.06042295  
## impulsiveness\_score 1.89547143  
## sens\_seeking\_score 0.21530081

Predicion for model 3

model.7\_fitted<-predict(model.7a, x.test1, type= "response")  
  
fitted.results.p2 <- ifelse(model.7\_fitted > 0.5,1,0)  
  
testing.model.2<-(as.numeric(test.data1$alc\_consumption))  
  
model.7\_Error <- mean(fitted.results.p2 != testing.model.2, na.rm=T)  
  
  
print(paste('Accuracy Model 3',1-model.7\_Error))

## [1] "Accuracy Model 3 0.769911504424779"

#### Question 2.

You should compare the performance of all three models within the test set and then decide which model you would choose as your final model. Provide justification for your choice

I would choose the elastice net or the logistic or traditional model becauese they had the same amount of accurancies when looking at the accurancy. The lasso has a lower accuracy and therefore should not be considered as a model to use.

#### Question 4.

What research questions could this analysis either a) directly address or b) indirectly help to address by providing information that could be used in subsequent analyses?

A research question that could be asked directly or indirectly to address by providing information that could be used in subsequent analyses would be doing a logistic regression to see if some of the features are associated with the alcohol consumptions .