Assignment 5: Predicting Current Alcohol Consumption from Behavioral Scores

R Setup Code:

### knitr::opts\_chunk$set(echo = TRUE)

### library(tidyverse)

### library(caret)

### library(glmnet)

### Creating and Comparing Three Different Models

#### 

#### Loading and Cleaning up “alcohol\_use” Dataset

R Code and Output:

set.seed(123)  
alcohol\_use = read\_csv("./alcohol\_use.csv")

Text

Description automatically generated

#Stripping off ID variable  
alcohol\_use<-alcohol\_use[,2:9]  
  
#Making sure variable types are correct  
str(alcohol\_use)

Text

Description automatically generated

#Changing outcome variable to factor and releveling to make "NotCurrentUse" as reference group  
alcohol\_use$alc\_consumption<-as.factor(alcohol\_use$alc\_consumption)  
alcohol\_use$alc\_consumption<-relevel(alcohol\_use$alc\_consumption, ref="NotCurrentUse")  
  
#Omit missing data  
alcohol\_use<-na.omit(alcohol\_use)

#### Partitioning Outcome Variable (alc\_consumption) into Training and Testing (70/30 Split)

R Code and Output:

set.seed(123)

train.indices<-createDataPartition(y=alcohol\_use$alc\_consumption,p=0.7,list=FALSE)  


#Training set (70%)  
train.data<-alcohol\_use[train.indices, ]

  
  
#Testing set (30%)  
test.data<-alcohol\_use[-train.indices, ] 

#### 

#### **Tuning and Comparing Performance of All Three Models Within Training Set Using Cross-Validation**

R Code and Output:

*\*Note: The numeric variables will be scaled and centered within the train functions.*

**Model 1: Model that chooses alpha and lambda via cross-validation using all of the features (Elastic Net Model)**

set.seed(123)  
  
cross.model<- train(  
 alc\_consumption ~., data = train.data, method = "glmnet",  
 trControl = trainControl("cv", number = 10), preProc=c("center", "scale"),  
 tuneLength=10  
 )  
  
cross.model$bestTune



confusionMatrix(cross.model)

Text

Description automatically generated with medium confidence

**Model 2: Model that uses all the features and traditional logistic regression**

set.seed(123)  
  
trad.model <- train(  
 alc\_consumption ~., data = train.data, method = "glm", family = "binomial",  
 trControl = trainControl("cv", number = 10), preProcess=c("center", "scale"))  
  
trad.model$bestTune



confusionMatrix(trad.model)

Text

Description automatically generated

**Model 3: Lasso Model using all of the features Tuning both alpha and lambda**

set.seed(123)  
  
#Creating grid to search lambda  
lambda<-10^seq(-3,3, length=100)  


#Fixing alpha = 1 and lambda = lambda  
lasso.model <- train(  
 alc\_consumption ~., data=train.data, method="glmnet", trControl=trainControl("cv", number=10), preProc=c("center", "scale"), tuneGrid=expand.grid(alpha=1, lambda=lambda)  
)  
  
lasso.model$bestTune



confusionMatrix(lasso.model)

Text

Description automatically generated with medium confidence

### 

### Which model would you choose as your final model? Justify your choice.

I have chosen the Elastic Net Model as my final model because it is the comprise of a ridge regression and lasso model. Given that the accuracy of both Model 1 (elastic net model) and Model 3 (lasso model) is the same at 86.14%, I will choose the Elastic Net Model because it allows me to keep the feature selection quality from the lasso penalty as well as the reduce overfitting while keeping all the features in the model from the ridge regression penalty.

### Applying Final Model to Test Set and Reporting Final Evaluation Metrics

**Model 3: Elastic Net Model Using All of the Features (model that chooses alpha and lambda via cross-validation using all of the features)**

R Code and Output:

set.seed(123)  
  
#Make predictions in test set  
test.outcome<-predict(cross.model, newdata=test.data)  


#Model Prediction Performance  
confusionMatrix(test.outcome, test.data$alc\_consumption, positive="CurrentUse")

Text

Description automatically generated

**5. 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? Limit this response to no more than 1 paragraph.**

A research question this analysis could indirectly help to address is does living in households where there is frequent drug and alcohol use predict children’s alcohol consumption as adults. There are different features that can determine a child’s personality trait. These features can be professionally recorded at a doctor’s visit as well as through informal surveys and questionnaires available online. If a researcher is trying to determine if a child’s living environment can predict a child’s future alcohol use, they can use this analysis, which can used in subsequent analyses.