Emily Murray

Feb. 25, 2019

BAN 502 – Week 5, Asmt #1

### Libraries & Data

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.5.2

## -- Attaching packages -------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.8  
## v tidyr 0.8.2 v stringr 1.3.1  
## v readr 1.3.1 v forcats 0.3.0

## Warning: package 'readr' was built under R version 3.5.2

## Warning: package 'dplyr' was built under R version 3.5.2

## -- Conflicts ----------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(caret)

## Warning: package 'caret' was built under R version 3.5.2

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

#install.packages("nnet")  
library(nnet)  
library(e1071)

## Warning: package 'e1071' was built under R version 3.5.2

parole = read\_csv("parole.csv")

## Parsed with column specification:  
## cols(  
## male = col\_double(),  
## race = col\_double(),  
## age = col\_double(),  
## state = col\_double(),  
## time.served = col\_double(),  
## max.sentence = col\_double(),  
## multiple.offenses = col\_double(),  
## crime = col\_double(),  
## violator = col\_double()  
## )

parole = parole %>%  
 mutate(male = as\_factor(as.character(male))) %>%  
 mutate(male = fct\_recode(male, "male" = "1", "female" = "0")) %>%  
 mutate(race = as\_factor(as.character(race))) %>%  
 mutate(race = fct\_recode(race, "white" = "1", "otherwise" = "2")) %>%  
 mutate(state = as\_factor(as.character(state))) %>%  
 mutate(state = fct\_recode(state, "kentucky" = "2", "louisiana" = "3", "virginia" = "4", "other" = "1")) %>%  
 mutate(crime = as\_factor(as.character(crime))) %>%  
 mutate(crime = fct\_recode(crime, "larceny" = "2", "drug" = "3", "driving" = "4", "other" = "1")) %>%  
 mutate(multiple.offenses = as\_factor(as.character(multiple.offenses))) %>%  
 mutate(multiple.offenses = fct\_recode(multiple.offenses, "multiple" = "1", "single" = "0")) %>%  
 mutate(violator = as\_factor(as.character(violator))) %>%  
 mutate(violator = fct\_recode(violator, "violated" = "1", "didnotviolate" = "0"))

### Task 2

set.seed(12345)   
train.rows = createDataPartition(y = parole$violator, p=0.7, list = FALSE)  
ptrain = parole[train.rows,]   
ptest = parole[-train.rows,]  
  
  
fitControla = trainControl(method = "cv",   
 number = 10)  
  
nnetGrida <- expand.grid(size = 12, decay = 0.1)  
  
  
set.seed(1234)  
nnetBasica = train(violator ~ .,   
 ptrain,  
 method = "nnet",  
 tuneGrid = nnetGrida,  
 trControl = fitControla,  
 verbose = FALSE,  
 trace = FALSE)

### Task 3

predNetBasica = predict(nnetBasica, ptrain)  
  
confusionMatrix(predNetBasica, ptrain$violator, positive = "violated")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction didnotviolate violated  
## didnotviolate 414 16  
## violated 4 39  
##   
## Accuracy : 0.9577   
## 95% CI : (0.9355, 0.974)  
## No Information Rate : 0.8837   
## P-Value [Acc > NIR] : 1.25e-08   
##   
## Kappa : 0.7727   
## Mcnemar's Test P-Value : 0.01391   
##   
## Sensitivity : 0.70909   
## Specificity : 0.99043   
## Pos Pred Value : 0.90698   
## Neg Pred Value : 0.96279   
## Prevalence : 0.11628   
## Detection Rate : 0.08245   
## Detection Prevalence : 0.09091   
## Balanced Accuracy : 0.84976   
##   
## 'Positive' Class : violated   
##

*Model looks pretty good. High accuracy of ~91.3, low p-value. It also has a high specificity, meaning it identifies true negatives well.*

### Task 4

start\_time = Sys.time()   
nnetGridb = expand.grid(size = seq(from = 2, to = 12, by = 1),   
 decay = seq(from = 0.1, to = 0.5, by = 0.1))  
set.seed(1234)  
fitControlb = trainControl(method = "cv",   
 number = 10)  
  
nnetBasicb = train(violator ~ .,   
 ptrain,  
 method = "nnet",  
 trControl = fitControlb,  
 tuneGrid = nnetGridb,  
 verbose = FALSE,  
 trace = FALSE)  
  
end\_time = Sys.time()  
end\_time-start\_time

## Time difference of 55.21537 secs

### Task 5

trainpredb = predict(nnetBasicb, ptrain)  
  
confusionMatrix(trainpredb, ptrain$violator, positive = "violated")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction didnotviolate violated  
## didnotviolate 408 33  
## violated 10 22  
##   
## Accuracy : 0.9091   
## 95% CI : (0.8795, 0.9334)  
## No Information Rate : 0.8837   
## P-Value [Acc > NIR] : 0.0459665   
##   
## Kappa : 0.4595   
## Mcnemar's Test P-Value : 0.0007937   
##   
## Sensitivity : 0.40000   
## Specificity : 0.97608   
## Pos Pred Value : 0.68750   
## Neg Pred Value : 0.92517   
## Prevalence : 0.11628   
## Detection Rate : 0.04651   
## Detection Prevalence : 0.06765   
## Balanced Accuracy : 0.68804   
##   
## 'Positive' Class : violated   
##

*Model still looks good on training set. Accuracy down slightly to about 90.9, p-value still below .05 threshold. Specificity is down slightly, but sensitivity is up considerably, meaning it identifies true positives better.*

### Task 6

predtesta = predict(nnetBasica, ptest)  
  
confusionMatrix(predtesta, ptest$violator, positive = "violated")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction didnotviolate violated  
## didnotviolate 174 17  
## violated 5 6  
##   
## Accuracy : 0.8911   
## 95% CI : (0.8398, 0.9305)  
## No Information Rate : 0.8861   
## P-Value [Acc > NIR] : 0.46721   
##   
## Kappa : 0.3015   
## Mcnemar's Test P-Value : 0.01902   
##   
## Sensitivity : 0.26087   
## Specificity : 0.97207   
## Pos Pred Value : 0.54545   
## Neg Pred Value : 0.91099   
## Prevalence : 0.11386   
## Detection Rate : 0.02970   
## Detection Prevalence : 0.05446   
## Balanced Accuracy : 0.61647   
##   
## 'Positive' Class : violated   
##

*Model is not great. Accuracy is not significantly better than with no information, and p-value is .47. Sensitivity is down.*

### Task 7

predtestb = predict(nnetBasicb, ptest)  
  
confusionMatrix(predtestb, ptest$violator, positive = "violated")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction didnotviolate violated  
## didnotviolate 175 19  
## violated 4 4  
##   
## Accuracy : 0.8861   
## 95% CI : (0.8341, 0.9264)  
## No Information Rate : 0.8861   
## P-Value [Acc > NIR] : 0.555251   
##   
## Kappa : 0.2117   
## Mcnemar's Test P-Value : 0.003509   
##   
## Sensitivity : 0.1739   
## Specificity : 0.9777   
## Pos Pred Value : 0.5000   
## Neg Pred Value : 0.9021   
## Prevalence : 0.1139   
## Detection Rate : 0.0198   
## Detection Prevalence : 0.0396   
## Balanced Accuracy : 0.5758   
##   
## 'Positive' Class : violated   
##

*Model is at its worst here. Accuracy is about the same as the NIR, and p-value is .56. Sensitivity is way down and it only correctly identified 4 out of 23 violators…*

### Task 8

\*Both models seem to have a problem with overfitting. It is definitely over-fitting more on the test data in the model from Task 4.