ML Final

Liz G

12/9/2021

Table of Contents

# Set Up

## Packages and Import

require(caret)

## Loading required package: caret

## Loading required package: ggplot2

## Loading required package: lattice

require(recipes)

## Loading required package: recipes

## Loading required package: dplyr

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

##   
## Attaching package: 'recipes'

## The following object is masked from 'package:stats':  
##   
## step

require(ranger)

## Loading required package: ranger

require(tidyverse)

## Loading required package: tidyverse

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble 3.1.5 v purrr 0.3.4  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.0.2 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x stringr::fixed() masks recipes::fixed()  
## x dplyr::lag() masks stats::lag()  
## x purrr::lift() masks caret::lift()

require(ModelMetrics)

## Loading required package: ModelMetrics

##   
## Attaching package: 'ModelMetrics'

## The following objects are masked from 'package:caret':  
##   
## confusionMatrix, precision, recall, sensitivity, specificity

## The following object is masked from 'package:base':  
##   
## kappa

require(here)

## Loading required package: here

## here() starts at C:/Users/epgle/Desktop/Fall 2021 Classes/EDLD 654 Machine Learning/Class Assignments/Final Project/ML-Final

require(finalfit)

## Loading required package: finalfit

require(Hmisc)

## Loading required package: Hmisc

## Loading required package: survival

##   
## Attaching package: 'survival'

## The following object is masked from 'package:caret':  
##   
## cluster

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

require(rsample)

## Loading required package: rsample

require(vip)

## Loading required package: vip

##   
## Attaching package: 'vip'

## The following object is masked from 'package:utils':  
##   
## vi

og\_data <- read.csv(here("data", "OPP\_PRO\_Clean.csv"), na.strings = "")  
  
og\_data <- og\_data %>% select(-starts\_with("x")) %>% #taking out unwanted variables  
 select(-matches("PSI\_total\_r|PSI\_total\_PR")) %>%  
 select(study\_id, PSI\_total\_clinical, any\_of(names(og\_data))) %>%  
 filter(!is.na(PSI\_total\_clinical))  
  
  
#taking out unwanted variables, putting in order  
  
head(og\_data)

## study\_id PSI\_total\_clinical TC\_age\_mos PC\_age AC\_age TC\_gender PC\_gender  
## 1 OP101 Yes 42 37 40 Male Female  
## 2 OP102 Yes 32 27 NA Female Female  
## 3 OP103 Yes 42 30 29 Male Female  
## 4 OP104 No 30 38 35 Male Male  
## 5 OP105 Yes 41 32 37 Male Female  
## 6 OP106 Yes 41 37 38 Male Female  
## PC\_reltoTC PC\_marital\_status PC\_partner\_inhome TC\_race\_Wh TC\_race\_Bl  
## 1 Bio Mom Married Yes 1 0  
## 2 Bio Mom Single No 1 0  
## 3 Bio Mom Married Yes 1 0  
## 4 Bio Dad Living together Yes 1 0  
## 5 Bio Mom Married Yes 1 0  
## 6 Foster Mom Married Yes 0 0  
## TC\_race\_His TC\_race\_As TC\_race\_Nat TC\_race\_Pac TC\_race\_Oth  
## 1 0 0 0 0 0  
## 2 0 0 0 0 0  
## 3 0 0 0 0 0  
## 4 0 0 0 0 0  
## 5 0 0 0 0 0  
## 6 0 0 1 0 0  
## PC\_yrs\_ed AC\_yrs\_ed  
## 1 HS grad-GED Specialized training  
## 2 HS grad-GED <NA>  
## 3 Partial high school Partial high school  
## 4 Partial college Partial college  
## 5 Partial college Partial college  
## 6 Standard college-university graduate Junior college-Associates degree  
## PC\_employment PC\_employ\_hours AC\_employment PC\_annual\_income  
## 1 Full time homemaker 0 Full time $80,000 to $89,999  
## 2 Unemployed 0 <NA> $5,000 to $9,999  
## 3 Full time homemaker 0 Full time $20,000 to $24,999  
## 4 Part time 24 Full time homemaker $10,000 to $14,999  
## 5 Full time homemaker 0 Part time $40,000 to $49,999  
## 6 Part time 20 Full time $80,000 to $89,999  
## gov\_assist PC\_fam\_learning PC\_fam\_mentalh TC\_diagnosis  
## 1 1 No No ASD  
## 2 1 No Yes Speech/Language Delay  
## 3 1 No Yes Social-Emotional Disorder  
## 4 1 Yes No ASD  
## 5 1 No No Speech/Language Delay  
## 6 1 Yes No Speech/Language Delay  
## TC\_diagnosis\_age TC\_SPED TC\_other\_services dyad\_adjust  
## 1 Two-years old (24-35 months) Yes Yes 104  
## 2 One-year old (12-23 months) Yes Yes 102  
## 3 Three-years old (36-47 months) Yes Yes 112  
## 4 One-year old (12-23 months) Yes Yes 98  
## 5 Three-years old (36-47 months) Yes Yes 84  
## 6 Two-years old (24-35 months) Yes Yes 139  
## fam\_support informal\_support formal\_support CES\_D CBCL\_int\_r CBCL\_ext\_r  
## 1 3 13 13 30 29 19  
## 2 8 6 20 19 32 46  
## 3 8 11 20 38 26 25  
## 4 9 8 25 17 21 17  
## 5 11 8 6 22 22 39  
## 6 23 23 23 1 25 28  
## VABS\_comm\_ss VABS\_dls\_ss VABS\_soc\_ss VABS\_mot\_ss PCBOS\_Aggr\_CU  
## 1 59 64 65 81 0  
## 2 69 78 80 69 0  
## 3 65 75 66 78 0  
## 4 66 87 68 100 0  
## 5 74 83 61 88 0  
## 6 69 81 85 91 0  
## PCBOS\_Aggr\_Play PCBOS\_Aggr\_SA PCBOS\_Crit\_CU PCBOS\_Crit\_Play PCBOS\_Crit\_SA  
## 1 0 0 0 0 0  
## 2 0 0 0 0 2  
## 3 0 0 0 0 0  
## 4 0 0 0 0 0  
## 5 0 0 0 0 4  
## 6 0 0 0 0 0  
## PCBOS\_CU\_Comm PCBOS\_CU\_Comp PCBOS\_DescComm\_CU PCBOS\_DescComm\_Play  
## 1 8 0 1 1  
## 2 31 2 0 5  
## 3 8 8 0 6  
## 4 7 0 0 2  
## 5 16 12 0 3  
## 6 1 1 1 7  
## PCBOS\_DescComm\_SA PCBOS\_Disr\_CU PCBOS\_Disr\_Play PCBOS\_Disr\_SA PCBOS\_InComm\_CU  
## 1 0 0 0 0 2  
## 2 0 0 3 0 4  
## 3 4 0 0 0 3  
## 4 3 0 0 0 1  
## 5 3 0 0 0 3  
## 6 2 0 0 0 0  
## PCBOS\_InComm\_Play PCBOS\_InComm\_SA PCBOS\_Intr\_CU PCBOS\_Intr\_Play PCBOS\_Intr\_SA  
## 1 7 4 1 4 2  
## 2 9 5 0 3 3  
## 3 10 1 0 0 1  
## 4 7 4 0 0 0  
## 5 4 0 1 1 5  
## 6 3 1 0 2 0  
## PCBOS\_IP\_CU PCBOS\_IP\_Play PCBOS\_IP\_SA PCBOS\_LackFT\_CU PCBOS\_LackFT\_Play  
## 1 0 15 5 1 10  
## 2 0 8 3 4 7  
## 3 0 12 2 3 1  
## 4 0 4 3 0 0  
## 5 0 4 1 3 0  
## 6 0 6 4 0 0  
## PCBOS\_LackFT\_SA PCBOS\_NegV\_CU PCBOS\_NegV\_Play PCBOS\_NegV\_SA PCBOS\_PhysAgg\_CU  
## 1 2 3 14 6 0  
## 2 4 4 0 0 0  
## 3 0 0 0 0 0  
## 4 1 3 0 0 0  
## 5 0 1 1 0 0  
## 6 1 0 0 0 0  
## PCBOS\_PhysAgg\_Play PCBOS\_PhysAgg\_SA PCBOS\_PosConsq\_CU PCBOS\_PosConsq\_Play  
## 1 0 0 0 0  
## 2 0 0 4 2  
## 3 0 0 0 0  
## 4 0 0 0 0  
## 5 0 0 0 0  
## 6 0 0 0 0  
## PCBOS\_PosConsq\_SA PCBOS\_PosV\_CU PCBOS\_PosV\_Play PCBOS\_PosV\_SA PCBOS\_Praise\_CU  
## 1 0 0 4 0 2  
## 2 0 3 17 2 3  
## 3 0 0 0 0 1  
## 4 0 0 0 0 0  
## 5 0 3 17 5 1  
## 6 0 4 19 6 3  
## PCBOS\_Praise\_Play PCBOS\_Praise\_SA TC\_sib\_age\_mean TC\_sib\_behavior\_ct  
## 1 3 0 7.000000 0  
## 2 6 8 4.666667 2  
## 3 3 5 7.000000 1  
## 4 7 3 15.000000 0  
## 5 2 5 8.500000 0  
## 6 11 2 9.000000 3  
## TC\_sib\_learning\_ct TC\_sib\_mentalh\_ct TC\_sib\_n\_ct TC\_sib\_yes\_ct  
## 1 1 0 1 1  
## 2 2 0 3 1  
## 3 0 1 2 1  
## 4 1 0 1 1  
## 5 1 0 2 1  
## 6 3 4 7 1

## Visualizing Missing Data

look <- ff\_glimpse(og\_data)  
  
look$Continuous %>% arrange(missing\_n)

## label var\_type n missing\_n missing\_percent  
## TC\_age\_mos TC\_age\_mos <int> 358 1 0.3  
## TC\_race\_Wh TC\_race\_Wh <int> 358 1 0.3  
## TC\_race\_Bl TC\_race\_Bl <int> 358 1 0.3  
## TC\_race\_His TC\_race\_His <int> 358 1 0.3  
## TC\_race\_As TC\_race\_As <int> 358 1 0.3  
## TC\_race\_Nat TC\_race\_Nat <int> 358 1 0.3  
## TC\_race\_Pac TC\_race\_Pac <int> 358 1 0.3  
## TC\_race\_Oth TC\_race\_Oth <int> 358 1 0.3  
## PC\_employ\_hours PC\_employ\_hours <int> 358 1 0.3  
## gov\_assist gov\_assist <int> 358 1 0.3  
## TC\_sib\_yes\_ct TC\_sib\_yes\_ct <int> 358 1 0.3  
## CBCL\_int\_r CBCL\_int\_r <int> 352 7 1.9  
## CBCL\_ext\_r CBCL\_ext\_r <int> 352 7 1.9  
## PC\_age PC\_age <int> 350 9 2.5  
## PCBOS\_Aggr\_CU PCBOS\_Aggr\_CU <int> 344 15 4.2  
## PCBOS\_Aggr\_Play PCBOS\_Aggr\_Play <int> 344 15 4.2  
## PCBOS\_Aggr\_SA PCBOS\_Aggr\_SA <int> 344 15 4.2  
## PCBOS\_Crit\_CU PCBOS\_Crit\_CU <int> 344 15 4.2  
## PCBOS\_Crit\_Play PCBOS\_Crit\_Play <int> 344 15 4.2  
## PCBOS\_Crit\_SA PCBOS\_Crit\_SA <int> 344 15 4.2  
## PCBOS\_CU\_Comm PCBOS\_CU\_Comm <int> 344 15 4.2  
## PCBOS\_CU\_Comp PCBOS\_CU\_Comp <int> 344 15 4.2  
## PCBOS\_DescComm\_CU PCBOS\_DescComm\_CU <int> 344 15 4.2  
## PCBOS\_DescComm\_Play PCBOS\_DescComm\_Play <int> 344 15 4.2  
## PCBOS\_DescComm\_SA PCBOS\_DescComm\_SA <int> 344 15 4.2  
## PCBOS\_Disr\_CU PCBOS\_Disr\_CU <int> 344 15 4.2  
## PCBOS\_Disr\_Play PCBOS\_Disr\_Play <int> 344 15 4.2  
## PCBOS\_Disr\_SA PCBOS\_Disr\_SA <int> 344 15 4.2  
## PCBOS\_InComm\_CU PCBOS\_InComm\_CU <int> 344 15 4.2  
## PCBOS\_InComm\_Play PCBOS\_InComm\_Play <int> 344 15 4.2  
## PCBOS\_InComm\_SA PCBOS\_InComm\_SA <int> 344 15 4.2  
## PCBOS\_Intr\_CU PCBOS\_Intr\_CU <int> 344 15 4.2  
## PCBOS\_Intr\_Play PCBOS\_Intr\_Play <int> 344 15 4.2  
## PCBOS\_Intr\_SA PCBOS\_Intr\_SA <int> 344 15 4.2  
## PCBOS\_IP\_CU PCBOS\_IP\_CU <int> 344 15 4.2  
## PCBOS\_IP\_Play PCBOS\_IP\_Play <int> 344 15 4.2  
## PCBOS\_IP\_SA PCBOS\_IP\_SA <int> 344 15 4.2  
## PCBOS\_LackFT\_CU PCBOS\_LackFT\_CU <int> 344 15 4.2  
## PCBOS\_LackFT\_Play PCBOS\_LackFT\_Play <int> 344 15 4.2  
## PCBOS\_LackFT\_SA PCBOS\_LackFT\_SA <int> 344 15 4.2  
## PCBOS\_NegV\_CU PCBOS\_NegV\_CU <int> 344 15 4.2  
## PCBOS\_NegV\_Play PCBOS\_NegV\_Play <int> 344 15 4.2  
## PCBOS\_NegV\_SA PCBOS\_NegV\_SA <int> 344 15 4.2  
## PCBOS\_PhysAgg\_CU PCBOS\_PhysAgg\_CU <int> 344 15 4.2  
## PCBOS\_PhysAgg\_Play PCBOS\_PhysAgg\_Play <int> 344 15 4.2  
## PCBOS\_PhysAgg\_SA PCBOS\_PhysAgg\_SA <int> 344 15 4.2  
## PCBOS\_PosConsq\_CU PCBOS\_PosConsq\_CU <int> 344 15 4.2  
## PCBOS\_PosConsq\_Play PCBOS\_PosConsq\_Play <int> 344 15 4.2  
## PCBOS\_PosConsq\_SA PCBOS\_PosConsq\_SA <int> 344 15 4.2  
## PCBOS\_PosV\_CU PCBOS\_PosV\_CU <int> 344 15 4.2  
## PCBOS\_PosV\_Play PCBOS\_PosV\_Play <int> 344 15 4.2  
## PCBOS\_PosV\_SA PCBOS\_PosV\_SA <int> 344 15 4.2  
## PCBOS\_Praise\_CU PCBOS\_Praise\_CU <int> 344 15 4.2  
## PCBOS\_Praise\_Play PCBOS\_Praise\_Play <int> 344 15 4.2  
## PCBOS\_Praise\_SA PCBOS\_Praise\_SA <int> 344 15 4.2  
## fam\_support fam\_support <int> 334 25 7.0  
## informal\_support informal\_support <int> 334 25 7.0  
## formal\_support formal\_support <int> 334 25 7.0  
## CES\_D CES\_D <dbl> 334 25 7.0  
## AC\_age AC\_age <int> 297 62 17.3  
## VABS\_comm\_ss VABS\_comm\_ss <int> 285 74 20.6  
## VABS\_dls\_ss VABS\_dls\_ss <int> 285 74 20.6  
## VABS\_soc\_ss VABS\_soc\_ss <int> 285 74 20.6  
## dyad\_adjust dyad\_adjust <int> 279 80 22.3  
## TC\_sib\_behavior\_ct TC\_sib\_behavior\_ct <int> 279 80 22.3  
## TC\_sib\_learning\_ct TC\_sib\_learning\_ct <int> 279 80 22.3  
## TC\_sib\_mentalh\_ct TC\_sib\_mentalh\_ct <int> 279 80 22.3  
## TC\_sib\_n\_ct TC\_sib\_n\_ct <int> 279 80 22.3  
## TC\_sib\_age\_mean TC\_sib\_age\_mean <dbl> 278 81 22.6  
## VABS\_mot\_ss VABS\_mot\_ss <int> 273 86 24.0  
## mean sd min quartile\_25 median quartile\_75 max  
## TC\_age\_mos 43.4 10.5 24.0 36.0 41.0 48.0 71.0  
## TC\_race\_Wh 0.7 0.5 0.0 0.0 1.0 1.0 1.0  
## TC\_race\_Bl 0.1 0.3 0.0 0.0 0.0 0.0 1.0  
## TC\_race\_His 0.4 0.5 0.0 0.0 0.0 1.0 1.0  
## TC\_race\_As 0.0 0.2 0.0 0.0 0.0 0.0 1.0  
## TC\_race\_Nat 0.1 0.2 0.0 0.0 0.0 0.0 1.0  
## TC\_race\_Pac 0.0 0.1 0.0 0.0 0.0 0.0 1.0  
## TC\_race\_Oth 0.0 0.2 0.0 0.0 0.0 0.0 1.0  
## PC\_employ\_hours 21.8 26.6 0.0 0.0 11.0 40.0 99.0  
## gov\_assist 0.6 0.5 0.0 0.0 1.0 1.0 1.0  
## TC\_sib\_yes\_ct 0.8 0.4 0.0 1.0 1.0 1.0 1.0  
## CBCL\_int\_r 36.8 25.7 0.0 11.0 36.0 59.0 100.0  
## CBCL\_ext\_r 42.0 25.0 0.0 19.0 42.5 64.0 100.0  
## PC\_age 34.6 7.9 19.0 29.0 34.0 39.0 67.0  
## PCBOS\_Aggr\_CU 0.0 0.3 0.0 0.0 0.0 0.0 3.0  
## PCBOS\_Aggr\_Play 0.0 0.3 0.0 0.0 0.0 0.0 3.0  
## PCBOS\_Aggr\_SA 0.0 0.2 0.0 0.0 0.0 0.0 2.0  
## PCBOS\_Crit\_CU 0.0 0.2 0.0 0.0 0.0 0.0 3.0  
## PCBOS\_Crit\_Play 0.1 0.4 0.0 0.0 0.0 0.0 4.0  
## PCBOS\_Crit\_SA 0.1 0.4 0.0 0.0 0.0 0.0 4.0  
## PCBOS\_CU\_Comm 20.1 12.8 0.0 11.0 18.0 27.0 68.0  
## PCBOS\_CU\_Comp 5.2 4.0 0.0 2.0 5.0 8.0 21.0  
## PCBOS\_DescComm\_CU 0.3 0.7 0.0 0.0 0.0 0.2 4.0  
## PCBOS\_DescComm\_Play 6.3 3.8 0.0 3.0 6.0 9.0 18.0  
## PCBOS\_DescComm\_SA 1.9 1.5 0.0 1.0 2.0 3.0 6.0  
## PCBOS\_Disr\_CU 0.2 0.6 0.0 0.0 0.0 0.0 4.0  
## PCBOS\_Disr\_Play 0.3 1.2 0.0 0.0 0.0 0.0 13.0  
## PCBOS\_Disr\_SA 0.1 0.6 0.0 0.0 0.0 0.0 5.0  
## PCBOS\_InComm\_CU 3.1 1.0 0.0 3.0 3.0 4.0 4.0  
## PCBOS\_InComm\_Play 6.7 4.5 0.0 3.0 6.0 9.2 19.0  
## PCBOS\_InComm\_SA 2.9 1.8 0.0 2.0 3.0 4.0 6.0  
## PCBOS\_Intr\_CU 0.5 0.9 0.0 0.0 0.0 1.0 4.0  
## PCBOS\_Intr\_Play 5.6 4.0 0.0 3.0 5.0 8.0 20.0  
## PCBOS\_Intr\_SA 3.0 1.8 0.0 2.0 3.0 5.0 6.0  
## PCBOS\_IP\_CU 0.1 0.3 0.0 0.0 0.0 0.0 3.0  
## PCBOS\_IP\_Play 3.8 3.3 0.0 1.0 3.0 6.0 19.0  
## PCBOS\_IP\_SA 1.5 1.4 0.0 0.0 1.0 2.0 6.0  
## PCBOS\_LackFT\_CU 2.9 1.2 0.0 2.0 3.0 4.0 4.0  
## PCBOS\_LackFT\_Play 6.7 4.7 0.0 3.0 6.0 10.0 19.0  
## PCBOS\_LackFT\_SA 2.8 1.9 0.0 1.0 3.0 5.0 6.0  
## PCBOS\_NegV\_CU 1.1 1.4 0.0 0.0 0.0 2.0 4.0  
## PCBOS\_NegV\_Play 1.1 2.7 0.0 0.0 0.0 1.0 20.0  
## PCBOS\_NegV\_SA 0.8 1.5 0.0 0.0 0.0 1.0 6.0  
## PCBOS\_PhysAgg\_CU 0.0 0.1 0.0 0.0 0.0 0.0 2.0  
## PCBOS\_PhysAgg\_Play 0.0 0.2 0.0 0.0 0.0 0.0 2.0  
## PCBOS\_PhysAgg\_SA 0.0 0.2 0.0 0.0 0.0 0.0 2.0  
## PCBOS\_PosConsq\_CU 0.2 0.7 0.0 0.0 0.0 0.0 4.0  
## PCBOS\_PosConsq\_Play 0.2 0.7 0.0 0.0 0.0 0.0 7.0  
## PCBOS\_PosConsq\_SA 0.0 0.3 0.0 0.0 0.0 0.0 4.0  
## PCBOS\_PosV\_CU 2.3 1.6 0.0 1.0 3.0 4.0 4.0  
## PCBOS\_PosV\_Play 14.7 6.7 0.0 12.0 18.0 20.0 20.0  
## PCBOS\_PosV\_SA 4.2 2.1 0.0 3.0 5.0 6.0 6.0  
## PCBOS\_Praise\_CU 2.8 3.2 0.0 0.0 2.0 4.0 15.0  
## PCBOS\_Praise\_Play 6.5 6.5 0.0 2.0 5.0 9.0 41.0  
## PCBOS\_Praise\_SA 6.2 4.6 0.0 2.0 5.0 9.0 24.0  
## fam\_support 12.8 6.1 0.0 9.0 12.0 17.0 30.0  
## informal\_support 7.3 5.6 0.0 3.0 6.0 10.0 26.0  
## formal\_support 13.9 6.4 0.0 9.0 13.0 19.0 30.0  
## CES\_D 13.4 10.4 0.0 6.0 10.5 18.0 48.0  
## AC\_age 37.6 8.7 21.0 31.0 36.0 42.0 67.0  
## VABS\_comm\_ss 75.8 13.8 17.0 67.0 76.0 84.0 143.0  
## VABS\_dls\_ss 81.2 14.9 35.0 71.0 81.0 91.0 143.0  
## VABS\_soc\_ss 80.2 12.6 40.0 74.0 79.0 88.0 152.0  
## dyad\_adjust 108.4 24.2 5.0 97.0 113.0 125.0 151.0  
## TC\_sib\_behavior\_ct 0.4 0.7 0.0 0.0 0.0 1.0 3.0  
## TC\_sib\_learning\_ct 0.3 0.6 0.0 0.0 0.0 0.0 3.0  
## TC\_sib\_mentalh\_ct 0.3 0.6 0.0 0.0 0.0 0.0 4.0  
## TC\_sib\_n\_ct 1.8 1.2 1.0 1.0 1.0 2.0 7.0  
## TC\_sib\_age\_mean 6.9 4.9 0.0 3.0 6.0 10.0 26.0  
## VABS\_mot\_ss 83.3 14.4 37.0 73.0 82.0 93.0 144.0

look$Categorical %>% arrange(missing\_n)

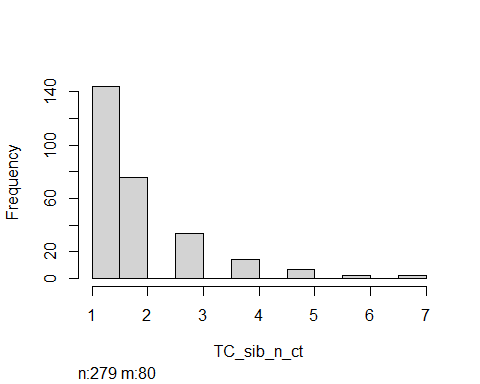
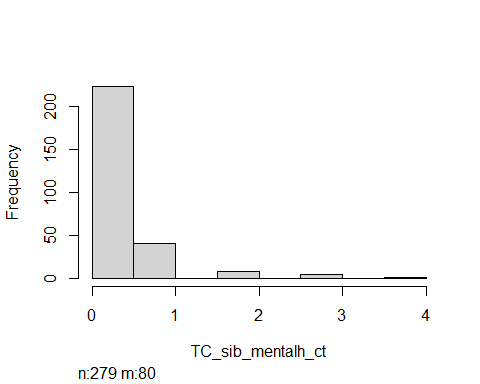
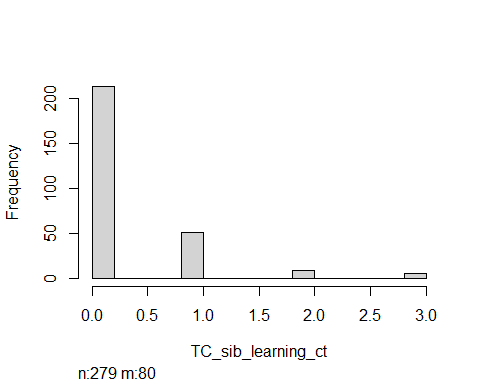
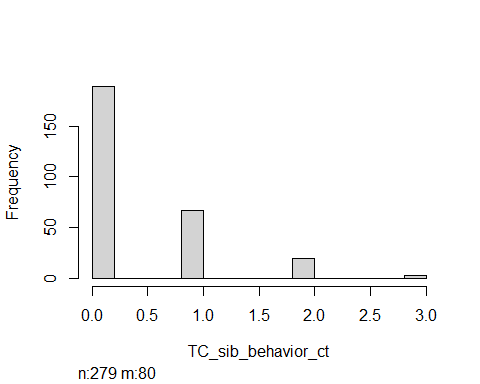
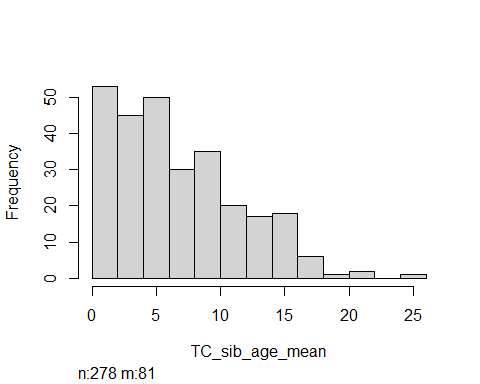
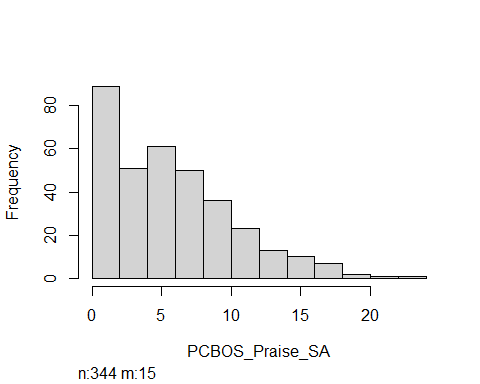
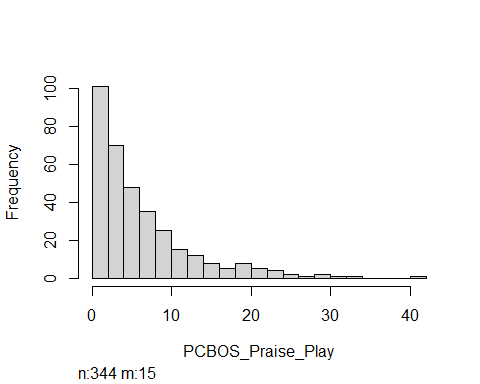
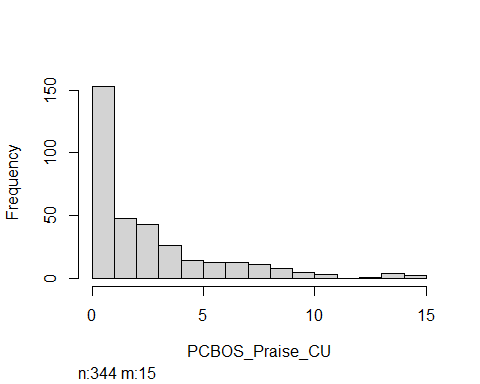
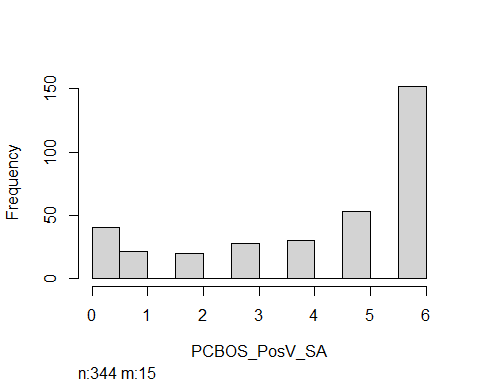
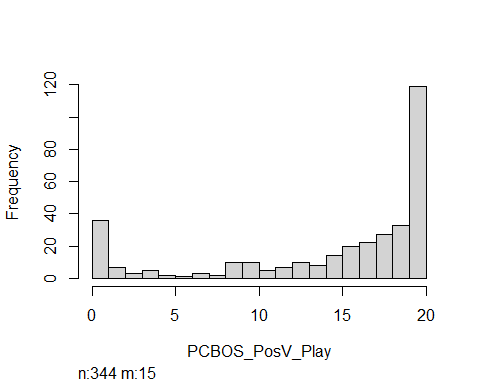
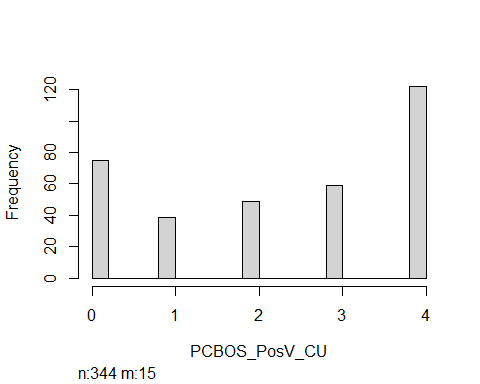
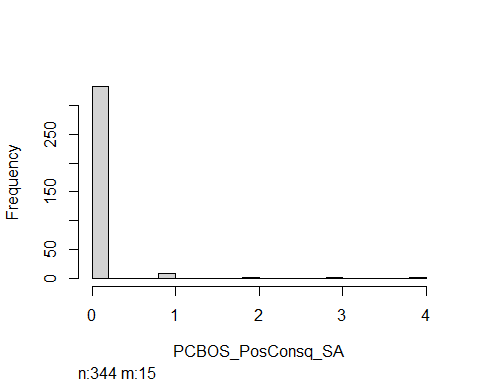
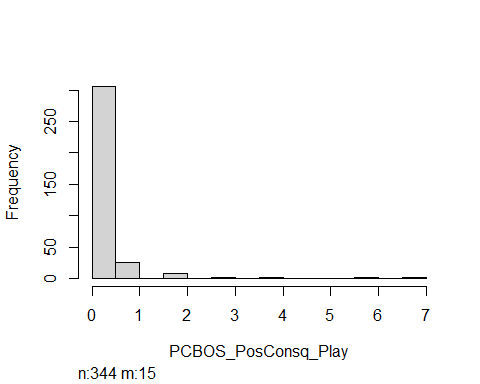
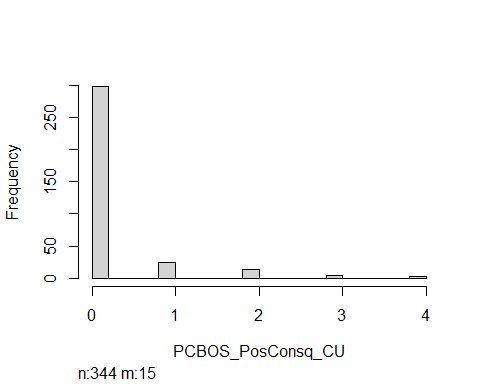
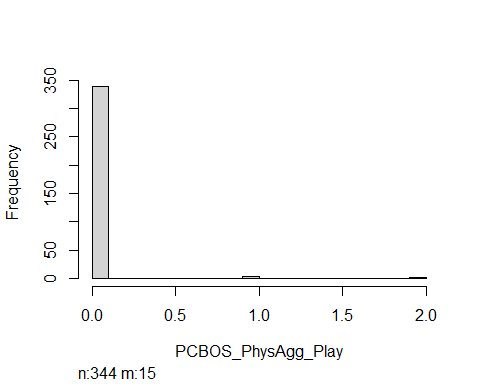
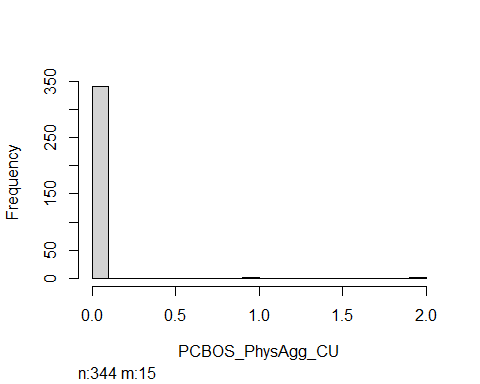
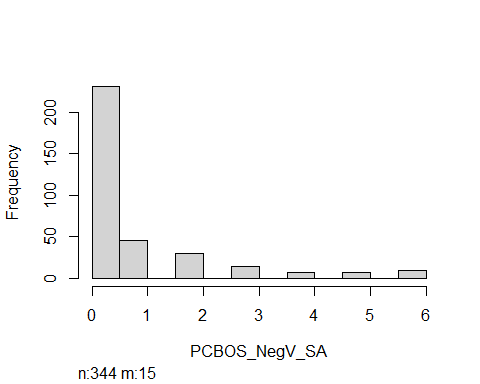
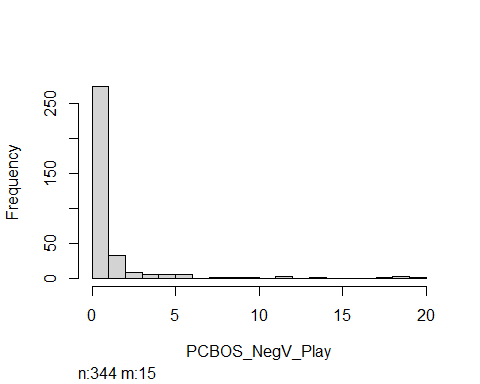
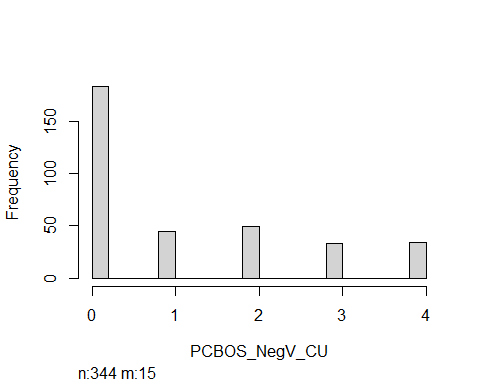
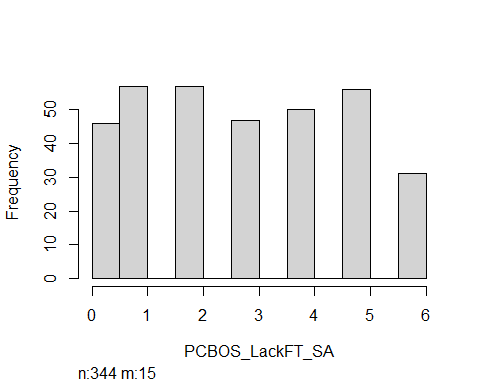
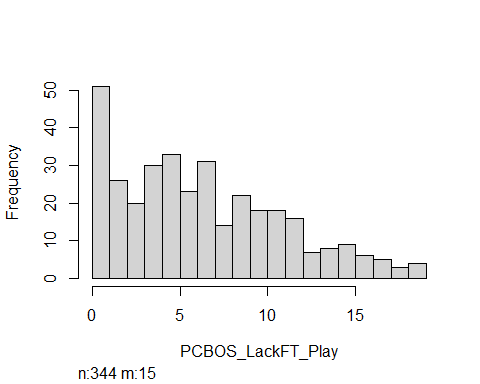
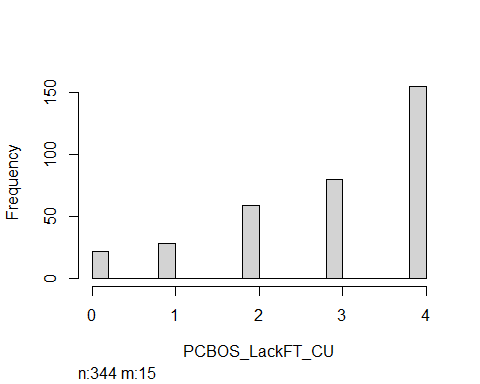
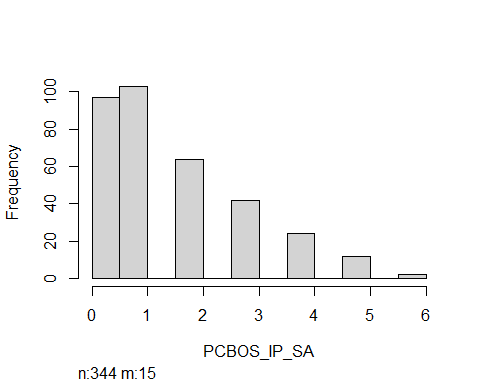
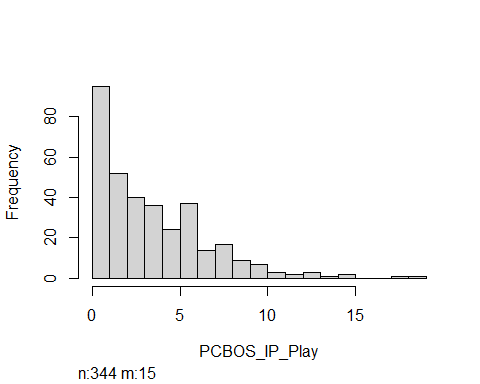
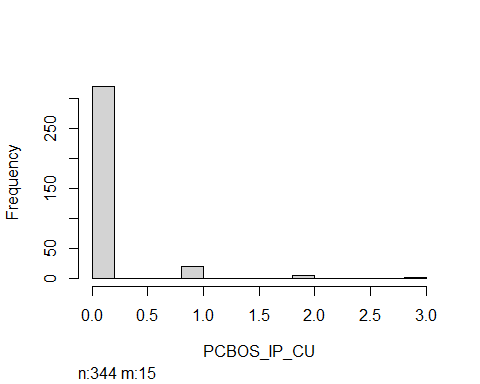
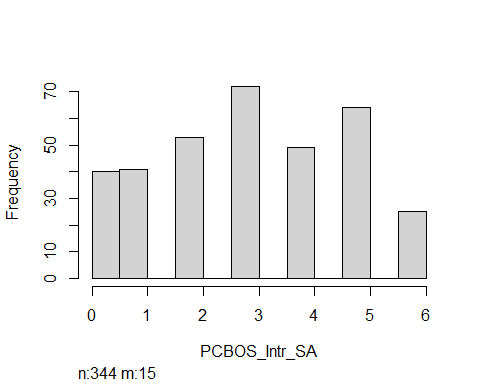
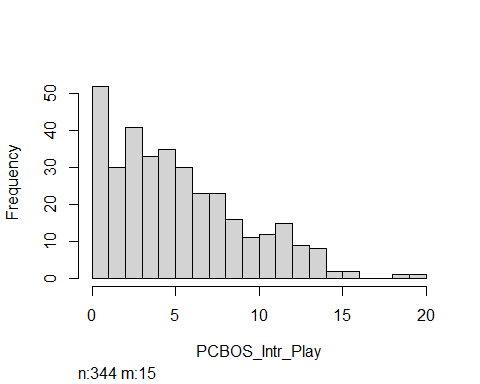
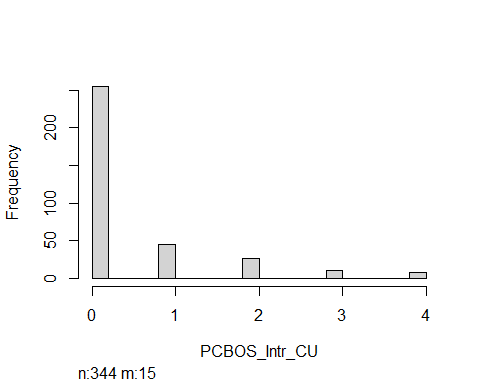
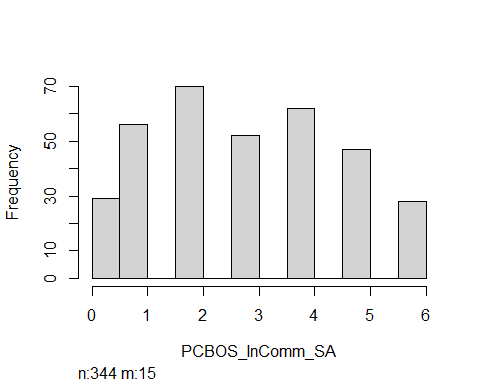
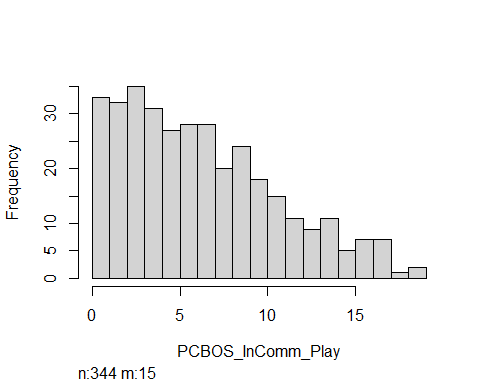
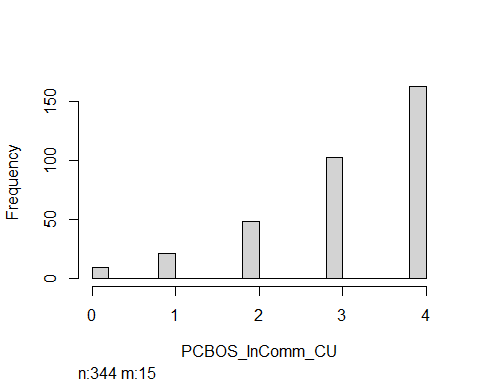
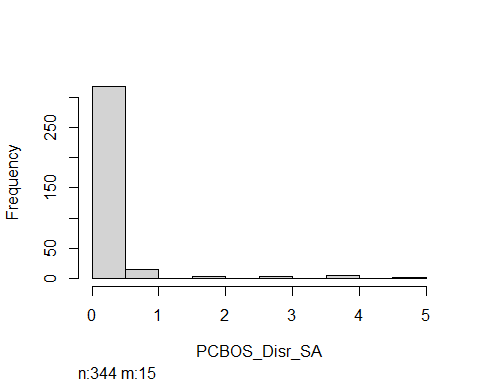
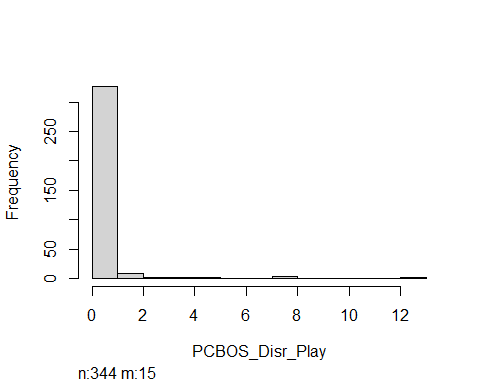
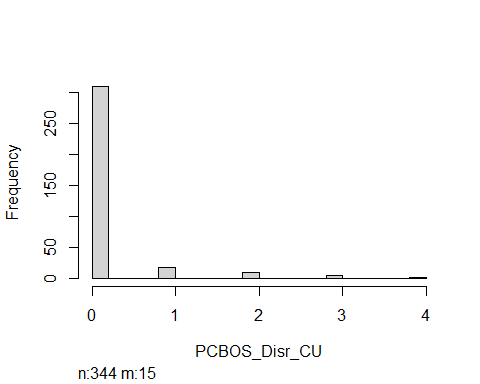
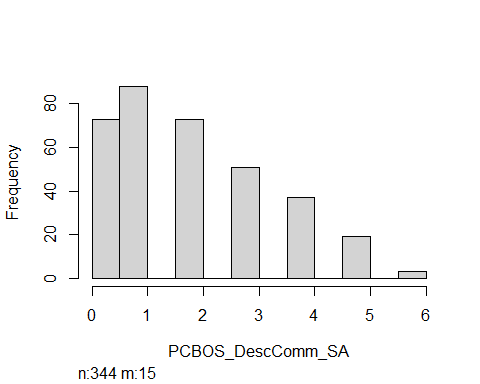
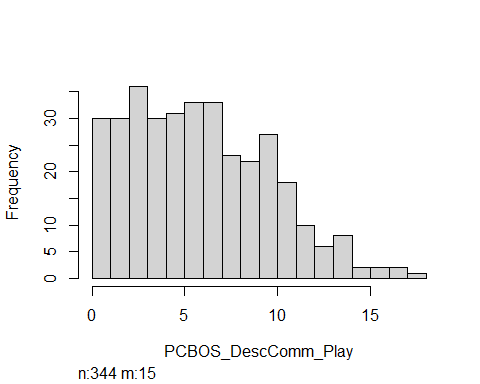
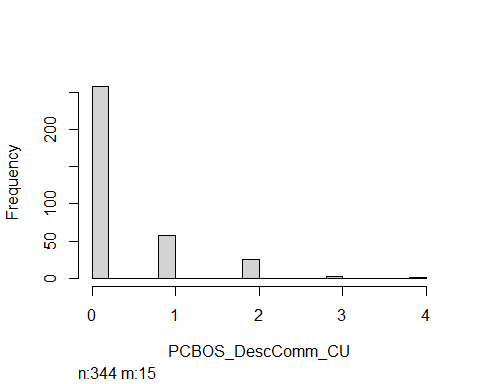
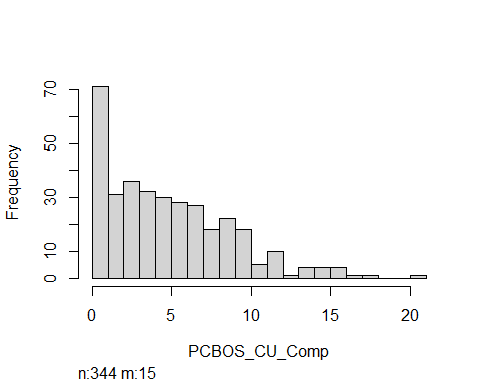
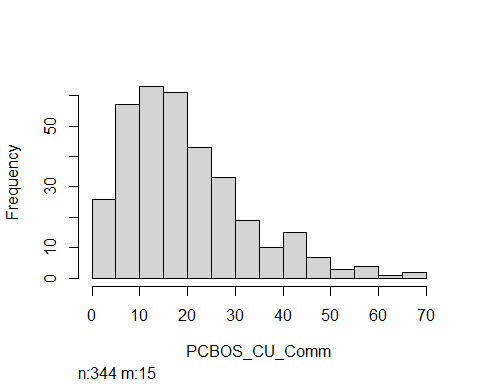
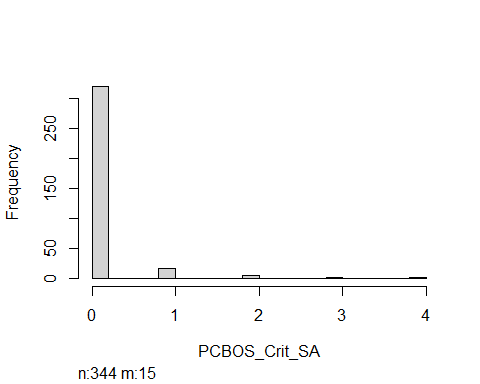
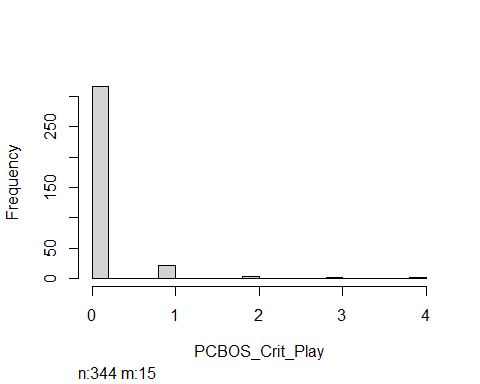
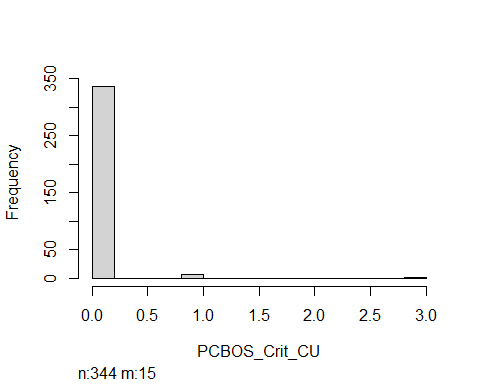
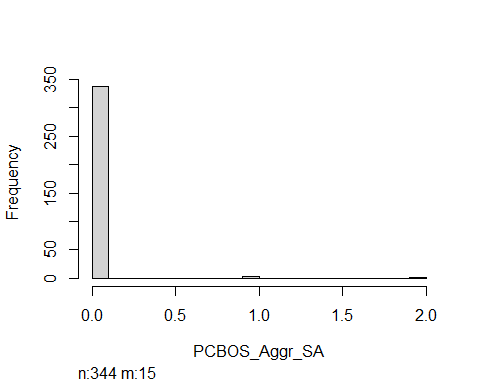
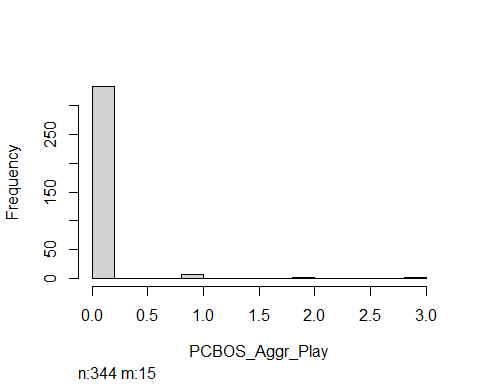
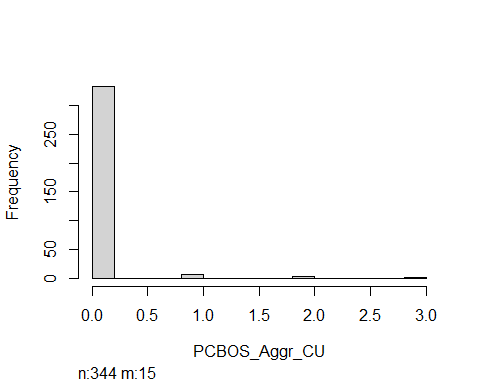
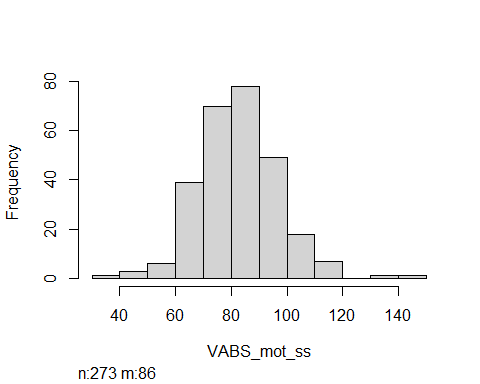
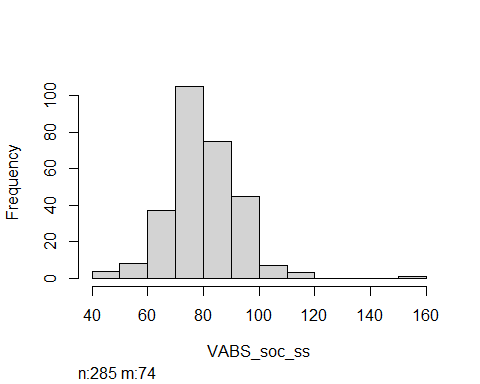
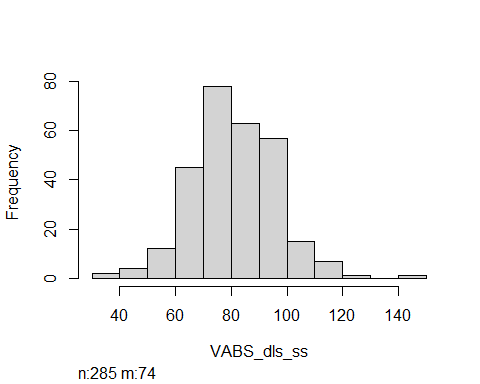
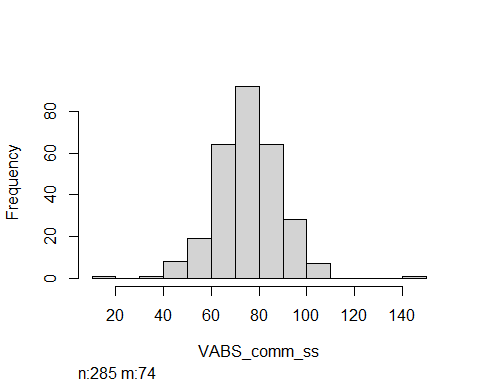
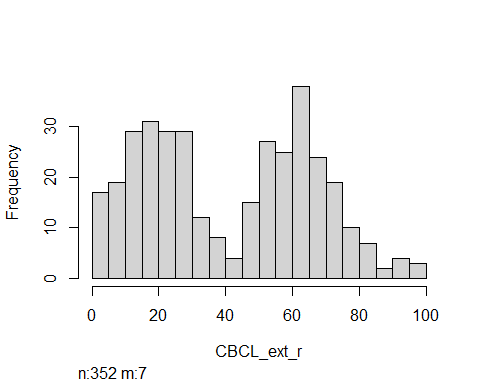
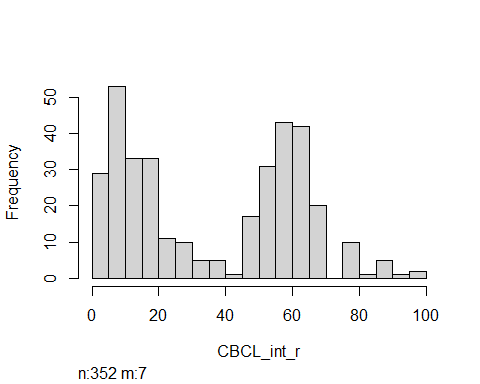
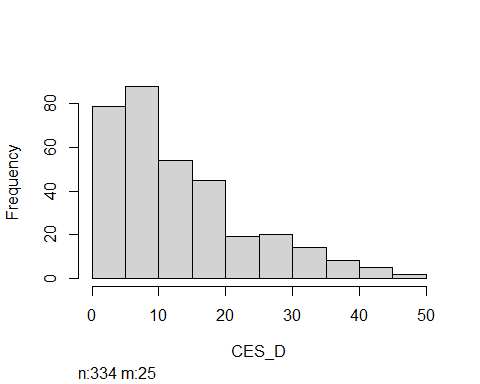
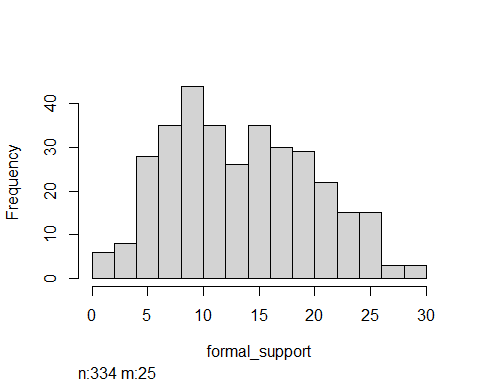
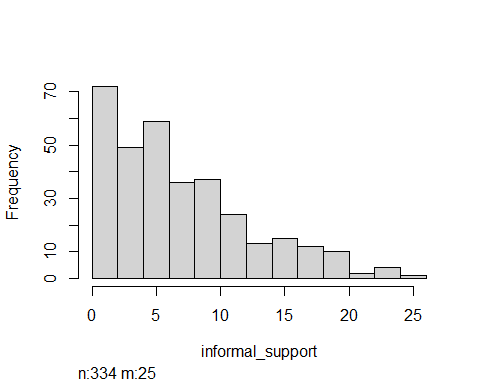
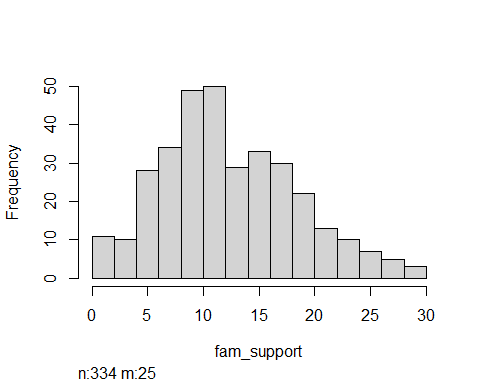
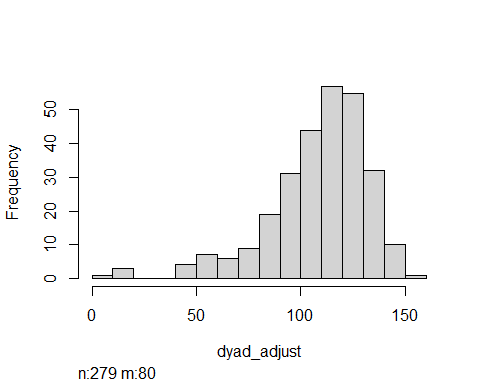
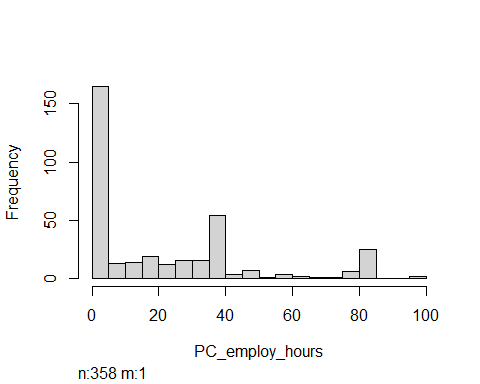
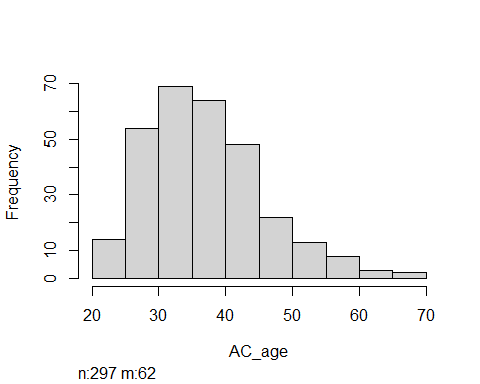
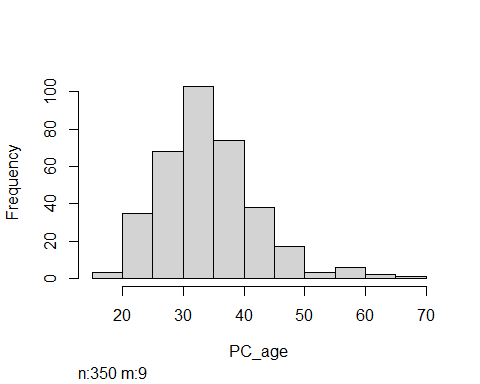
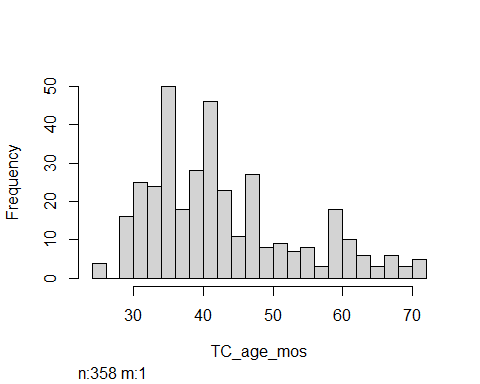
## label var\_type n missing\_n missing\_percent  
## study\_id study\_id <chr> 359 0 0.0  
## PSI\_total\_clinical PSI\_total\_clinical <chr> 359 0 0.0  
## TC\_gender TC\_gender <chr> 358 1 0.3  
## PC\_gender PC\_gender <chr> 358 1 0.3  
## PC\_reltoTC PC\_reltoTC <chr> 358 1 0.3  
## PC\_marital\_status PC\_marital\_status <chr> 358 1 0.3  
## PC\_partner\_inhome PC\_partner\_inhome <chr> 358 1 0.3  
## PC\_yrs\_ed PC\_yrs\_ed <chr> 358 1 0.3  
## PC\_employment PC\_employment <chr> 358 1 0.3  
## PC\_fam\_learning PC\_fam\_learning <chr> 358 1 0.3  
## PC\_fam\_mentalh PC\_fam\_mentalh <chr> 358 1 0.3  
## PC\_annual\_income PC\_annual\_income <chr> 357 2 0.6  
## TC\_other\_services TC\_other\_services <chr> 357 2 0.6  
## TC\_diagnosis TC\_diagnosis <chr> 356 3 0.8  
## TC\_SPED TC\_SPED <chr> 350 9 2.5  
## TC\_diagnosis\_age TC\_diagnosis\_age <chr> 346 13 3.6  
## AC\_yrs\_ed AC\_yrs\_ed <chr> 301 58 16.2  
## AC\_employment AC\_employment <chr> 300 59 16.4  
## levels\_n levels levels\_count levels\_percent  
## study\_id 359 - - -  
## PSI\_total\_clinical 2 - - -  
## TC\_gender 2 - - -  
## PC\_gender 2 - - -  
## PC\_reltoTC 9 - - -  
## PC\_marital\_status 6 - - -  
## PC\_partner\_inhome 2 - - -  
## PC\_yrs\_ed 16 - - -  
## PC\_employment 10 - - -  
## PC\_fam\_learning 2 - - -  
## PC\_fam\_mentalh 2 - - -  
## PC\_annual\_income 13 - - -  
## TC\_other\_services 2 - - -  
## TC\_diagnosis 15 - - -  
## TC\_SPED 4 - - -  
## TC\_diagnosis\_age 6 - - -  
## AC\_yrs\_ed 15 - - -  
## AC\_employment 9 - - -

#Examine levels for categorical data  
look$Categorical %>% select(label, n, levels\_n)

## label n levels\_n  
## study\_id study\_id 359 359  
## PSI\_total\_clinical PSI\_total\_clinical 359 2  
## TC\_gender TC\_gender 358 2  
## PC\_gender PC\_gender 358 2  
## PC\_reltoTC PC\_reltoTC 358 9  
## PC\_marital\_status PC\_marital\_status 358 6  
## PC\_partner\_inhome PC\_partner\_inhome 358 2  
## PC\_yrs\_ed PC\_yrs\_ed 358 16  
## AC\_yrs\_ed AC\_yrs\_ed 301 15  
## PC\_employment PC\_employment 358 10  
## AC\_employment AC\_employment 300 9  
## PC\_annual\_income PC\_annual\_income 357 13  
## PC\_fam\_learning PC\_fam\_learning 358 2  
## PC\_fam\_mentalh PC\_fam\_mentalh 358 2  
## TC\_diagnosis TC\_diagnosis 356 15  
## TC\_diagnosis\_age TC\_diagnosis\_age 346 6  
## TC\_SPED TC\_SPED 350 4  
## TC\_other\_services TC\_other\_services 357 2

#look @ distributions of numeric data  
cont\_labels <- look$Continuous$label  
  
for (i in cont\_labels) {  
 hist(og\_data %>% select(i))  
}

## Note: Using an external vector in selections is ambiguous.  
## i Use `all\_of(i)` instead of `i` to silence this message.  
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.  
## This message is displayed once per session.



Preprocessing Steps - 1) Get rid of rows with no outcome data (moved out of blueprint for error testing) 2) Missing variable column for - AC Age, TC\_sib (NA is meaningful) 3) Make yrs ed into numeric variable Make income into numeric variable Make age @ dx into numeric variable 4) For factor variables w/ high number of levels, make “other” 5) Dummy code categorical variables, and drop OG 6) Multiple imputation is reasonable for most variables 7) For variables where NA is meaningful (= absence of AC or sib, set remaining values to 0 for regression models)

## Making a Blueprint

outcome <- "PSI\_total\_clinical"  
cat2num <- names(select(og\_data, matches("yrs\_ed|diagnosis\_age|income"))) #variables that will be recoded to be numeric  
categorical <- names(select(og\_data, where(is.character) & !(matches("id|PSI"))))  
categorical <- categorical[!categorical %in% cat2num]  
numeric <- c(names(select(og\_data, where(is.numeric))), cat2num)  
already\_dum <- names(select(og\_data, matches("TC\_race|gov\_assist|TC\_sib\_yes\_ct"))) #some variables already "dummy" coded  
true\_numeric <- numeric[!numeric %in% already\_dum]  
na\_true <- names(select(og\_data, matches("AC|TC\_sib", ignore.case = FALSE))) #don't impute data for alternate caregiver or subling variables  
impute\_vars <- true\_numeric[!true\_numeric %in% na\_true]  
  
stress\_blueprint <- recipe(x = og\_data,   
 vars = names(og\_data),  
 roles = c("ID", "outcome", rep("predictor", 86))) %>%   
 #step\_filter(!is.na(PSI\_total\_clinical), skip = FALSE) %>% #causing error   
 step\_indicate\_na(matches("AC\_age", "TC\_sib\_yes\_ct")) %>%  
 step\_other(matches("PC\_reltoTC|employment|TC\_diagnosis|PC\_marital\_status|TC\_SPED")) %>%  
 #making yrs ed numeric  
 step\_mutate(across(matches("yrs\_ed"), ~case\_when(  
 str\_detect(., "No formal") ~ 0,  
 str\_detect(., "7th grade") ~ 7,   
 str\_detect(., "Junior high") ~ 8,  
 str\_detect(., "Partial high school") ~ 9,  
 str\_detect(., "GED") ~ 12,  
 str\_detect(., "Partial college") ~ 13,  
 str\_detect(., "Specialized") ~ 14,  
 str\_detect(., "Associates") ~ 14,  
 str\_detect(., "university grad") ~ 16,  
 str\_detect(., "Grad prof") ~ 18))) %>%  
 #making dx age numeric  
 step\_mutate(across(matches("diagnosis\_age"), ~case\_when(  
 str\_detect(., "At birth") ~ 6,  
 str\_detect(., "One") ~ 18,   
 str\_detect(., "Two") ~ 30,  
 str\_detect(., "Three") ~ 42,  
 str\_detect(., "Four") ~ 54,  
 str\_detect(., "Five") ~ 66))) %>%  
 #making annual income numeric  
 step\_mutate(PC\_annual\_income =   
 ifelse(str\_detect(PC\_annual\_income, "less"), 2500,  
 ifelse(str\_detect(PC\_annual\_income, "14,999"), 12500,  
 ifelse(str\_detect(PC\_annual\_income, "19,999"), 17500,  
 ifelse(str\_detect(PC\_annual\_income, "24,999"), 22500,  
 ifelse(str\_detect(PC\_annual\_income, "29,999"), 27500,  
 ifelse(str\_detect(PC\_annual\_income, "39,999"), 35000,  
 ifelse(str\_detect(PC\_annual\_income, "49,999"), 45000,  
 ifelse(str\_detect(PC\_annual\_income, "59,999"), 55000,  
 ifelse(str\_detect(PC\_annual\_income, "69,999"), 65000,  
 ifelse(str\_detect(PC\_annual\_income, "79,999"), 75000,  
 ifelse(str\_detect(PC\_annual\_income, "89,999"), 85000,  
 ifelse(str\_detect(PC\_annual\_income, "more"), 95000,  
 ifelse(is.na(PC\_annual\_income), NA, 7500)))))))))))))) %>% #5,000 and $9,999 in all others  
 step\_impute\_mode(all\_of(categorical)) %>%  
 step\_zv(all\_of(numeric)) %>%  
 step\_dummy(all\_of(categorical)) %>%  
 step\_mutate(across(matches("AC|TC\_sib", ignore.case = FALSE),   
 ~replace\_na(., 0))) %>%  
 step\_impute\_knn(all\_of(impute\_vars), all\_of(already\_dum)) %>%  
 step\_normalize(all\_of(true\_numeric))

prepare <- prep(stress\_blueprint, training = og\_data)  
stress\_data <- bake(prepare, new\_data = og\_data)  
  
head(stress\_data)

## # A tibble: 6 x 101  
## study\_id PSI\_total\_clinical TC\_age\_mos PC\_age AC\_age TC\_race\_Wh TC\_race\_Bl  
## <fct> <fct> <dbl> <dbl> <dbl> <int> <int>  
## 1 OP101 Yes -0.131 0.303 0.548 1 0  
## 2 OP102 Yes -1.08 -0.968 -1.91 1 0  
## 3 OP103 Yes -0.131 -0.587 -0.127 1 0  
## 4 OP104 No -1.27 0.430 0.241 1 0  
## 5 OP105 Yes -0.226 -0.332 0.364 1 0  
## 6 OP106 Yes -0.226 0.303 0.425 0 0  
## # ... with 94 more variables: TC\_race\_His <int>, TC\_race\_As <int>,  
## # TC\_race\_Nat <int>, TC\_race\_Pac <int>, TC\_race\_Oth <int>, PC\_yrs\_ed <dbl>,  
## # AC\_yrs\_ed <dbl>, PC\_employ\_hours <dbl>, PC\_annual\_income <dbl>,  
## # gov\_assist <int>, TC\_diagnosis\_age <dbl>, dyad\_adjust <dbl>,  
## # fam\_support <dbl>, informal\_support <dbl>, formal\_support <dbl>,  
## # CES\_D <dbl>, CBCL\_int\_r <dbl>, CBCL\_ext\_r <dbl>, VABS\_comm\_ss <dbl>,  
## # VABS\_dls\_ss <dbl>, VABS\_soc\_ss <dbl>, VABS\_mot\_ss <dbl>, ...

#Check for missingness  
look2 <- ff\_glimpse(stress\_data)  
  
look2$Continuous %>% arrange(desc(missing\_n)) %>% filter(missing\_n > 0)

## [1] label var\_type n missing\_n   
## [5] missing\_percent mean sd min   
## [9] quartile\_25 median quartile\_75 max   
## <0 rows> (or 0-length row.names)

look2$Categorical %>% arrange(missing\_n) %>% filter(missing\_n > 0)

## [1] label var\_type n missing\_n   
## [5] missing\_percent levels\_n levels levels\_count   
## [9] levels\_percent   
## <0 rows> (or 0-length row.names)

## Test and Training Set

set.seed(121221)  
 split <- initial\_split(og\_data, prop = .75)  
 stress\_train <- training(split)  
 stress\_test <- testing(split)  
   
prepare <- prep(stress\_blueprint,   
 training = stress\_train)  
  
baked\_train <- bake(prepare, new\_data = stress\_train)  
baked\_test <- bake(prepare, new\_data = stress\_test)

## Functions

##Making a crossfold function  
  
crossfold\_log <- function(training\_data, folds){  
   
 #shuffle data  
 traning\_data <- training\_data[sample(nrow(training\_data)),]  
   
  
 # Create 10 folds with equal size  
  
 N\_folds = cut(seq(1,nrow(training\_data)),breaks= folds,labels=FALSE)  
   
 # Create the list for each fold   
   
 my.indices <- vector('list',folds)  
 for(i in 1:folds){  
 my.indices[[i]] <- which(N\_folds!=i)  
 }  
   
 #cross validation settings  
   
 cv <- trainControl(method = "cv",  
 index = my.indices,  
 classProbs = TRUE,  
 summaryFunction = mnLogLoss)  
   
   
return(cv)  
}  
  
#Making an accuracy function   
  
accuracy <- function(observed\_vector, predicted\_vector){  
tab <- table(predicted\_vector,  
 observed\_vector,  
 dnn = c('Predicted','Observed'))  
  
  
  
tn <- tab[1,1]  
tp <- tab[2,2]  
fp <- tab[2,1]  
fn <- tab[1,2]  
  
acc <- (tp + tn)/(tp+tn+fp+fn)  
  
return(acc)  
}

# Model Building

## Model 1

Ridge Regression

cf <- crossfold\_log(stress\_train, 10) #error occurs for both stress\_train and baked\_train  
   
folds <- vfold\_cv(stress\_train, 10)  
  
cv <- trainControl(method = "cv",  
 number = 10,  
 classProbs = TRUE,  
 summaryFunction = mnLogLoss)  
  
  
  
ridge\_grid <- data.frame(alpha = 0, lambda = c(seq(.01, 5, .1)))  
  
  
  
mod\_1 <- caret::train(stress\_blueprint,  
 data = stress\_train,   
 method = "glmnet",  
 family = 'binomial',  
 metric = 'logLoss',  
 trControl = cf,  
 tuneGrid = ridge\_grid)  
mod\_1$bestTune  
plot(mod\_1)

I get an error - “number of rows of result is not a multiple of vector length (arg 1)”… It’s possible rows are not lining up correctly either w/ the blueprint or cross validation settings. blueprint seems to be working to “bake” data… This goes away when filtering step is applied outside of blueprint

When filter step applied prior to blueprint, get errors for applying blueprint within model..

## Model Eval Metrics and Variable Importance

predicted\_test <- predict(mod\_1, stress\_test) %>%   
 as.numeric() %>% -1 #subtracting 1 b/c factor levels come out as 1/2 rather than 0/1  
observed\_test <- stress\_test$PSI\_total\_clinical %>% as.numeric()  
  
  
#LogLoss ACC ROC Sensitivity Specificity PPV  
  
#LogLoss  
LL <- logLoss(observed\_test, predicted\_test)  
   
#AUC  
AUC <- auc(observed\_test, predicted\_test)  
  
#Accuracy  
ACC <- accuracy(observed\_test, predicted\_test)  
  
#Sensitivity  
TPR <- tpr(observed\_test, predicted\_test, cutoff = .5)  
  
#Specificity  
TNR <- tnr(observed\_test, predicted\_test, cutoff = .5)  
  
#PPV  
PRE <- precision(observed\_test, predicted\_test, cutoff = .5)  
  
mod\_1\_stats <- c("Ridge Regression", LL, AUC, TPR, TNR, PRE)  
  
#Looking @ VIP for 20 feats  
  
vip(mod\_1,   
 num\_features = 20,   
 geom = "point") +   
 theme\_bw()

## Model 2

Random Forests

cf <- crossfold\_log(stress\_train, 10)  
  
#tune trees   
#Can do similar tuning for max depth  
  
mod\_tune <- vector('list',200)  
   
 for(i in 1:200){  
   
 mod\_tune[[i]] <- caret::train(blueprint\_recidivism,  
 data = recidivism\_tr,  
 method = 'ranger',  
 trControl = cv,  
 tuneGrid = grid,  
 metric = 'logLoss',  
 num.trees = i,  
 max.depth = 60)  
 }  
  
  
logLoss\_ <- c()  
  
for(i in 1:200){  
   
 logLoss\_[i] = mod\_tune[[i]]$results$logLoss  
   
}  
  
  
ggplot()+  
 geom\_line(aes(x=1:200,y=logLoss\_))+  
 xlab('Number of Tree Models')+  
 ylab('Negative LogLoss')+  
 ylim(c(0,12))+  
 theme\_bw()

Only tune N predictors

# Grid settings  
grid <- expand.grid(mtry = 30,  
 splitrule='gini',  
 min.node.size=2)  
  
  
mod\_2 <- caret::train(stress\_blueprint,  
 data = stress\_train,  
 method = 'ranger',  
 trControl = cf,  
 tuneGrid = grid,  
 num.trees = 500,  
 max.depth = 10)

When filter in blueprint - Similar error occurs here, x Existing data has 27 rows. x Assigned data has 28 rows.

It seems something is potentially going badly when making the cv setting - perhaps because blueprint uses some filter function?

With filter not in blueprint: Currently get errors – Warning in gower\_work(x = x, y = y, pair\_x = pair\_x, pair\_y = pair\_y, n = n, : skipping variable with zero or non-finite range Warning: model fit failed for Resample09: mtry=30, splitrule=gini, min.node.size=2 Error : Can’t subset columns that don’t exist. x Column PCBOS\_PhysAgg\_SA doesn’t exist.

Error in { : task 1 failed - “$ operator is invalid for atomic vectors”

## Model 2 Performance

predicted\_test <- predict(mod\_2, stress\_test) %>%   
 as.numeric() %>% -1 #subtracting 1 b/c factor levels come out as 1/2 rather than 0/1  
observed\_test <- stress\_test$PSI\_total\_clinical %>% as.numeric()  
  
  
#LogLoss ACC ROC Sensitivity Specificity PPV  
  
#LogLoss  
LL <- logLoss(observed\_test, predicted\_test)  
   
#AUC  
AUC <- auc(observed\_test, predicted\_test)  
  
#Accuracy  
ACC <- accuracy(observed\_test, predicted\_test)  
  
#Sensitivity  
TPR <- tpr(observed\_test, predicted\_test, cutoff = .5)  
  
#Specificity  
TNR <- tnr(observed\_test, predicted\_test, cutoff = .5)  
  
#PPV  
PRE <- precision(observed\_test, predicted\_test, cutoff = .5)  
  
mod\_2\_stats <- c("Ridge Regression", LL, AUC, TPR, TNR, PRE)  
  
#Looking @ VIP for 20 feats  
  
vip(mod\_2,   
 num\_features = 20,   
 geom = "point") +   
 theme\_bw()

## Code Graveyard

#step\_mutate(PC\_reltoTC =   
 #ifelse(str\_detect(PC\_reltoTC, "Grandma|Grandpa|Aunt|Uncle"), "Kinship",   
 #ifelse(str\_detect(PC\_reltoTC, "Bio"), "Bio Parent",   
 #ifelse(str\_detect(PC\_reltoTC, "Adoptive|Partner|Step"), "Adoptive",   
 #ifelse(str\_detect(PC\_reltoTC, "Foster"), "Foster", NA)))))  
 #taken out for now, but could re-implement if need to get more specific