Mod4Assign2

Natasha Minott

2/13/2021

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.2 ✓ purrr 0.3.4  
## ✓ tibble 3.0.4 ✓ dplyr 1.0.2  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ forcats 0.5.0

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

library(tidymodels)

## ── Attaching packages ────────────────────────────────────── tidymodels 0.1.2 ──

## ✓ broom 0.7.2 ✓ recipes 0.1.15  
## ✓ dials 0.0.9 ✓ rsample 0.0.8   
## ✓ infer 0.5.4 ✓ tune 0.1.2   
## ✓ modeldata 0.1.0 ✓ workflows 0.2.1   
## ✓ parsnip 0.1.5 ✓ yardstick 0.0.7

## ── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
## x scales::discard() masks purrr::discard()  
## x dplyr::filter() masks stats::filter()  
## x recipes::fixed() masks stringr::fixed()  
## x dplyr::lag() masks stats::lag()  
## x yardstick::spec() masks readr::spec()  
## x recipes::step() masks stats::step()

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

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

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

library( gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(vip)

##   
## Attaching package: 'vip'

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

library(ranger)

drug<- read\_csv("~/Desktop/502 Predictive Analytics/MODULE 4/drug\_data-1.csv")

##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## .default = col\_character(),  
## Column1 = col\_double(),  
## Column2 = col\_double(),  
## Column3 = col\_double(),  
## Column4 = col\_double(),  
## Column5 = col\_double(),  
## Column6 = col\_double(),  
## Column7 = col\_double(),  
## Column8 = col\_double(),  
## Column9 = col\_double(),  
## Column10 = col\_double(),  
## Column11 = col\_double(),  
## Column12 = col\_double(),  
## Column13 = col\_double()  
## )  
## ℹ Use `spec()` for the full column specifications.

names(drug) =c("ID", "Age", "Gender", "Education", "Country", "Ethnicity","Nscore", "Escore", "Oscore", "Ascore", "Cscore", "Impulsive","SS", "Alcohol", "Amphet", "Amyl", "Benzos", "Caff", "Cannabis","Choc", "Coke", "Crack", "Ecstasy", "Heroin", "Ketamine", "Legalh","LSD", "Meth", "Mushrooms", "Nicotine", "Semer", "VSA")  
#str(drug)

drug[drug=="CL0"] = "No"  
drug[drug=="CL1"] = "No"  
drug[drug=="CL2"] = "Yes"  
drug[drug=="CL3"] = "Yes"  
drug[drug=="CL4"] = "Yes"  
drug[drug=="CL5"] = "Yes"  
drug[drug=="CL6"] = "Yes"

drug\_clean = drug%>% mutate\_at(vars(Age:Ethnicity),funs(as\_factor))%>%mutate(Age =factor(Age, labels =c("18\_24", "25\_34", "35\_44","45\_54", "55\_64", "65\_")))%>%mutate(Gender =factor(Gender, labels =c("Male", "Female")))%>%mutate(Education =factor(Education, labels =c("Under16", "At16", "At17", "At18", "SomeCollege","ProfessionalCert", "Bachelors", "Masters", "Doctorate")))%>%mutate(Country =factor(Country,labels =c("USA", "NewZealand", "Other", "Australia","Ireland","Canada","UK")))%>%mutate(Ethnicity =factor(Ethnicity,labels =c("Black", "Asian", "White", "White/Black", "Other","White/Asian", "Black/Asian")))%>%  
 mutate\_at(vars(Alcohol:VSA),funs(as\_factor))%>%select(-ID)

## Warning: `funs()` is deprecated as of dplyr 0.8.0.  
## Please use a list of either functions or lambdas:   
##   
## # Simple named list:   
## list(mean = mean, median = median)  
##   
## # Auto named with `tibble::lst()`:   
## tibble::lst(mean, median)  
##   
## # Using lambdas  
## list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_warnings()` to see where this warning was generated.

str(drug\_clean)

## tibble [1,885 × 31] (S3: tbl\_df/tbl/data.frame)  
## $ Age : Factor w/ 6 levels "18\_24","25\_34",..: 3 2 3 1 3 6 4 3 3 5 ...  
## $ Gender : Factor w/ 2 levels "Male","Female": 2 1 1 2 2 2 1 1 2 1 ...  
## $ Education: Factor w/ 9 levels "Under16","At16",..: 6 9 6 8 9 4 8 2 6 8 ...  
## $ Country : Factor w/ 7 levels "USA","NewZealand",..: 7 7 7 7 7 6 1 7 6 7 ...  
## $ Ethnicity: Factor w/ 7 levels "Black","Asian",..: 6 3 3 3 3 3 3 3 3 3 ...  
## $ Nscore : num [1:1885] 0.313 -0.678 -0.467 -0.149 0.735 ...  
## $ Escore : num [1:1885] -0.575 1.939 0.805 -0.806 -1.633 ...  
## $ Oscore : num [1:1885] -0.5833 1.4353 -0.8473 -0.0193 -0.4517 ...  
## $ Ascore : num [1:1885] -0.917 0.761 -1.621 0.59 -0.302 ...  
## $ Cscore : num [1:1885] -0.00665 -0.14277 -1.0145 0.58489 1.30612 ...  
## $ Impulsive: num [1:1885] -0.217 -0.711 -1.38 -1.38 -0.217 ...  
## $ SS : num [1:1885] -1.181 -0.216 0.401 -1.181 -0.216 ...  
## $ Alcohol : Factor w/ 2 levels "Yes","No": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Amphet : Factor w/ 2 levels "Yes","No": 1 1 2 2 2 2 2 2 2 2 ...  
## $ Amyl : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 1 ...  
## $ Benzos : Factor w/ 2 levels "Yes","No": 1 2 2 1 2 2 2 2 2 2 ...  
## $ Caff : Factor w/ 2 levels "Yes","No": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Cannabis : Factor w/ 2 levels "No","Yes": 1 2 2 2 2 1 1 1 1 1 ...  
## $ Choc : Factor w/ 2 levels "Yes","No": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Coke : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 1 1 1 1 1 ...  
## $ Crack : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Ecstasy : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 1 ...  
## $ Heroin : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Ketamine : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 1 1 1 1 1 ...  
## $ Legalh : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ LSD : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 1 ...  
## $ Meth : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 1 ...  
## $ Mushrooms: Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 1 1 1 ...  
## $ Nicotine : Factor w/ 2 levels "Yes","No": 1 1 2 1 1 1 1 2 1 1 ...  
## $ Semer : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ VSA : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...

drug\_clean = drug\_clean %>% select(!(Alcohol:Mushrooms)) %>% select(!(Semer:VSA))  
names(drug\_clean)

## [1] "Age" "Gender" "Education" "Country" "Ethnicity" "Nscore"   
## [7] "Escore" "Oscore" "Ascore" "Cscore" "Impulsive" "SS"   
## [13] "Nicotine"

## Task 1

There is not any missingness within this data set.

str(drug\_clean)

## tibble [1,885 × 13] (S3: tbl\_df/tbl/data.frame)  
## $ Age : Factor w/ 6 levels "18\_24","25\_34",..: 3 2 3 1 3 6 4 3 3 5 ...  
## $ Gender : Factor w/ 2 levels "Male","Female": 2 1 1 2 2 2 1 1 2 1 ...  
## $ Education: Factor w/ 9 levels "Under16","At16",..: 6 9 6 8 9 4 8 2 6 8 ...  
## $ Country : Factor w/ 7 levels "USA","NewZealand",..: 7 7 7 7 7 6 1 7 6 7 ...  
## $ Ethnicity: Factor w/ 7 levels "Black","Asian",..: 6 3 3 3 3 3 3 3 3 3 ...  
## $ Nscore : num [1:1885] 0.313 -0.678 -0.467 -0.149 0.735 ...  
## $ Escore : num [1:1885] -0.575 1.939 0.805 -0.806 -1.633 ...  
## $ Oscore : num [1:1885] -0.5833 1.4353 -0.8473 -0.0193 -0.4517 ...  
## $ Ascore : num [1:1885] -0.917 0.761 -1.621 0.59 -0.302 ...  
## $ Cscore : num [1:1885] -0.00665 -0.14277 -1.0145 0.58489 1.30612 ...  
## $ Impulsive: num [1:1885] -0.217 -0.711 -1.38 -1.38 -0.217 ...  
## $ SS : num [1:1885] -1.181 -0.216 0.401 -1.181 -0.216 ...  
## $ Nicotine : Factor w/ 2 levels "Yes","No": 1 1 2 1 1 1 1 2 1 1 ...

summary(drug\_clean)

## Age Gender Education Country   
## 18\_24:643 Male :943 SomeCollege :506 USA : 557   
## 25\_34:481 Female:942 Bachelors :480 NewZealand: 5   
## 35\_44:356 Masters :283 Other : 118   
## 45\_54:294 ProfessionalCert:270 Australia : 54   
## 55\_64: 93 At18 :100 Ireland : 20   
## 65\_ : 18 At16 : 99 Canada : 87   
## (Other) :147 UK :1044   
## Ethnicity Nscore Escore Oscore   
## Black : 33 Min. :-3.464360 Min. :-3.273930 Min. :-3.273930   
## Asian : 26 1st Qu.:-0.678250 1st Qu.:-0.695090 1st Qu.:-0.717270   
## White :1720 Median : 0.042570 Median : 0.003320 Median :-0.019280   
## White/Black: 20 Mean : 0.000047 Mean :-0.000163 Mean :-0.000534   
## Other : 63 3rd Qu.: 0.629670 3rd Qu.: 0.637790 3rd Qu.: 0.723300   
## White/Asian: 20 Max. : 3.273930 Max. : 3.273930 Max. : 2.901610   
## Black/Asian: 3   
## Ascore Cscore Impulsive   
## Min. :-3.464360 Min. :-3.464360 Min. :-2.555240   
## 1st Qu.:-0.606330 1st Qu.:-0.652530 1st Qu.:-0.711260   
## Median :-0.017290 Median :-0.006650 Median :-0.217120   
## Mean :-0.000245 Mean :-0.000386 Mean : 0.007216   
## 3rd Qu.: 0.760960 3rd Qu.: 0.584890 3rd Qu.: 0.529750   
## Max. : 3.464360 Max. : 3.464360 Max. : 2.901610   
##   
## SS Nicotine   
## Min. :-2.078480 Yes:1264   
## 1st Qu.:-0.525930 No : 621   
## Median : 0.079870   
## Mean :-0.003292   
## 3rd Qu.: 0.765400   
## Max. : 1.921730   
##

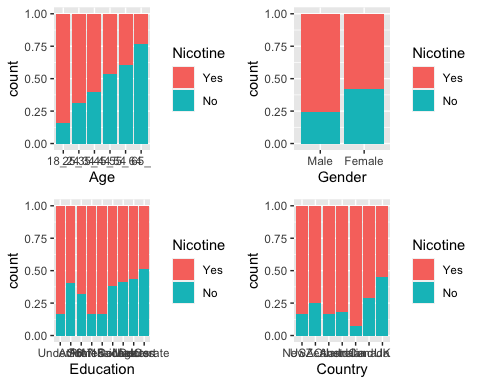
## Task 2

set.seed(1234)   
drug\_split = initial\_split(drug\_clean, prop = 0.7, strata = Nicotine)   
train = training(drug\_split)   
test = testing(drug\_split)

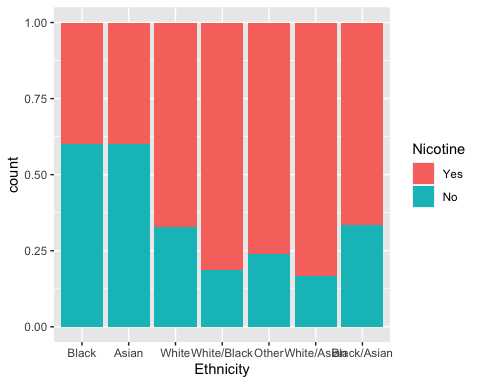
## Task 3

As age increases, nicotine use decreases. More males use Nicotine than females. As a person’s education increases, they are less likely to use Nicotine. The Uk has the least amount of Nicotine users. Black and Asians use Nicotine more than any other race. People with high N & O scores are more likely to use Nicotine. Higher Ascores are less likely to use it.

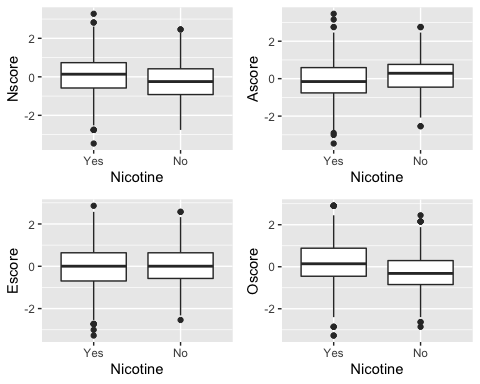
p1 =ggplot(train, aes(x=Age, fill=Nicotine)) +  
geom\_bar(position ="fill")  
p2 = ggplot(train, aes(x =Gender, fill = Nicotine)) +   
geom\_bar(position = "fill")  
p3 = ggplot(train, aes(x= Education, fill = Nicotine)) +  
geom\_bar(position = "fill")  
p4 = ggplot(train, aes(x= Country, fill = Nicotine)) +  
geom\_bar(position = "fill")  
grid.arrange(p1,p2,p3,p4)



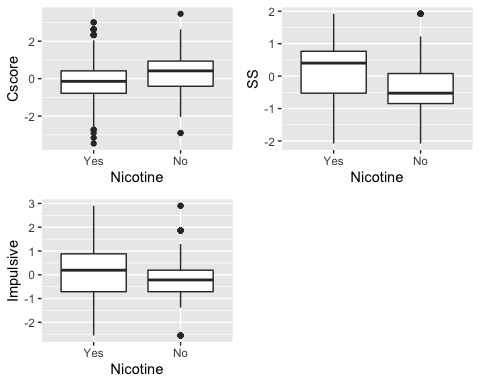
p1 = ggplot(train, aes(x =Ethnicity, fill = Nicotine)) +  
geom\_bar(position = "fill")  
grid.arrange(p1)



p1 = ggplot(train, aes(x = Nicotine, y= Nscore)) + geom\_boxplot()  
p2 = ggplot(train, aes(x = Nicotine, y = Ascore)) + geom\_boxplot()  
p3 = ggplot(train, aes(x = Nicotine, y = Escore)) + geom\_boxplot()  
p4 = ggplot(train, aes(x = Nicotine, y = Oscore)) + geom\_boxplot()  
grid.arrange(p1,p2,p3,p4, ncol = 2)



p1 = ggplot(train, aes(x = Nicotine, y = Cscore)) + geom\_boxplot()  
p2 = ggplot(train, aes(x = Nicotine, y = SS)) + geom\_boxplot()  
p3 = ggplot(drug\_clean, aes(x = Nicotine, y = Impulsive)) + geom\_boxplot()  
grid.arrange(p1,p2,p3, ncol = 2)



## Task 4

set.seed(123)  
rf\_folds = vfold\_cv(train, v = 5)  
  
  
drug\_recipe = recipe(Nicotine ~., train) %>%  
 step\_dummy(all\_nominal(), -all\_outcomes())  
rf\_model = rand\_forest(mtry = tune(), min\_n = tune(), trees = 100) %>%   
 set\_engine("ranger", importance = "permutation") %>%   
 set\_mode("classification")  
  
drug\_wflow =   
 workflow() %>%   
 add\_model(rf\_model) %>%   
 add\_recipe(drug\_recipe)  
  
rf\_grid = grid\_regular(  
 mtry(range = c(2, 18)),   
 min\_n(range = c(5, 20)),   
 levels = 10  
)  
  
set.seed(123)  
rf\_res\_tuned = tune\_grid(  
 drug\_wflow,  
 resamples = rf\_folds,  
 grid = rf\_grid  
)

##   
## Attaching package: 'rlang'

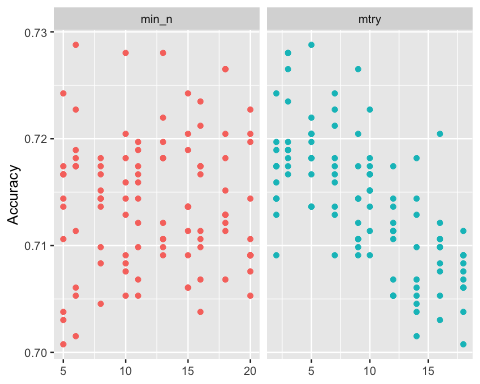
## The following objects are masked from 'package:purrr':  
##   
## %@%, as\_function, flatten, flatten\_chr, flatten\_dbl, flatten\_int,  
## flatten\_lgl, flatten\_raw, invoke, list\_along, modify, prepend,  
## splice

##   
## Attaching package: 'vctrs'

## The following object is masked from 'package:dplyr':  
##   
## data\_frame

## The following object is masked from 'package:tibble':  
##   
## data\_frame

rf\_res\_tuned %>%  
 collect\_metrics() %>%  
 filter(.metric == "accuracy") %>%  
 select(mean, min\_n, mtry) %>%  
 pivot\_longer(min\_n:mtry,  
 values\_to = "value",  
 names\_to = "parameter"  
 ) %>%  
 ggplot(aes(value, mean, color = parameter)) +  
 geom\_point(show.legend = FALSE) +  
 facet\_wrap(~parameter, scales = "free\_x") +  
 labs(x = NULL, y = "Accuracy")



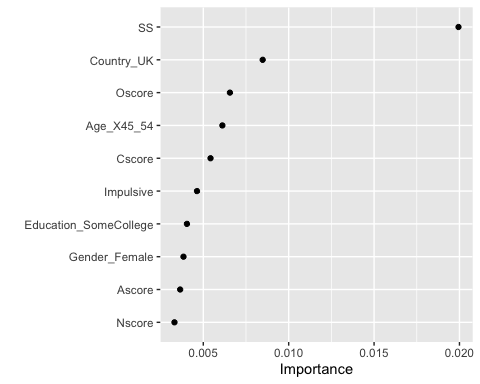
## Task 5

Within this model, SS seems to be the most important variable followed by the country UK.

best\_rf = select\_best(rf\_res\_tuned, "accuracy")  
  
final\_rf =finalize\_workflow(  
 drug\_wflow,  
 best\_rf  
)  
  
final\_rf

## ══ Workflow ════════════════════════════════════════════════════════════════════  
## Preprocessor: Recipe  
## Model: rand\_forest()  
##   
## ── Preprocessor ────────────────────────────────────────────────────────────────  
## 1 Recipe Step  
##   
## ● step\_dummy()  
##   
## ── Model ───────────────────────────────────────────────────────────────────────  
## Random Forest Model Specification (classification)  
##   
## Main Arguments:  
## mtry = 5  
## trees = 100  
## min\_n = 6  
##   
## Engine-Specific Arguments:  
## importance = permutation  
##   
## Computational engine: ranger

final\_rf\_fit = fit(final\_rf, train)  
final\_rf\_fit %>% pull\_workflow\_fit() %>% vip(geom = "point")



## Task 6

Also this model works good on the training set, there is a reduction in accuracy when it is performed on the testing set.

trainpredrf = predict(final\_rf\_fit, train)  
head(trainpredrf)

## # A tibble: 6 x 1  
## .pred\_class  
## <fct>   
## 1 Yes   
## 2 Yes   
## 3 No   
## 4 Yes   
## 5 Yes   
## 6 Yes

confusionMatrix(trainpredrf$.pred\_class, train$Nicotine,   
 positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 882 32  
## No 3 403  
##   
## Accuracy : 0.9735   
## 95% CI : (0.9633, 0.9815)  
## No Information Rate : 0.6705   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.939   
##   
## Mcnemar's Test P-Value : 2.214e-06   
##   
## Sensitivity : 0.9966   
## Specificity : 0.9264   
## Pos Pred Value : 0.9650   
## Neg Pred Value : 0.9926   
## Prevalence : 0.6705   
## Detection Rate : 0.6682   
## Detection Prevalence : 0.6924   
## Balanced Accuracy : 0.9615   
##   
## 'Positive' Class : Yes   
##

testpredrf = predict(final\_rf\_fit, test)  
head(testpredrf)

## # A tibble: 6 x 1  
## .pred\_class  
## <fct>   
## 1 Yes   
## 2 Yes   
## 3 No   
## 4 Yes   
## 5 No   
## 6 Yes

confusionMatrix(testpredrf$.pred\_class, test$Nicotine,   
 positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 326 113  
## No 53 73  
##   
## Accuracy : 0.7062   
## 95% CI : (0.6667, 0.7435)  
## No Information Rate : 0.6708   
## P-Value [Acc > NIR] : 0.03945   
##   
## Kappa : 0.2752   
##   
## Mcnemar's Test P-Value : 4.666e-06   
##   
## Sensitivity : 0.8602   
## Specificity : 0.3925   
## Pos Pred Value : 0.7426   
## Neg Pred Value : 0.5794   
## Prevalence : 0.6708   
## Detection Rate : 0.5770   
## Detection Prevalence : 0.7770   
## Balanced Accuracy : 0.6263   
##   
## 'Positive' Class : Yes   
##

## Task 7

This model could be used to target a specific demographic that may be likely to smoke. Understanding this demographic could help create commericals that would encourage these peoeple to stop smoking. My biggest concern for this model is it’s inconsistency due to overfitting