Week 5 AHA: Random Forests

2024-02-21

Install packages

```r
# Install packages
install.packages(
  c("ranger", "randomForest", "randomForestExplainer")
)
```

Load packages (and set seed)

```r
# Load packages
library(randomForest); library(randomForestExplainer)
library(caret); library(ranger)
library(ggplot2); library(ggpubr)
set.seed(42)
```

Load data

```r
# Load data
schizotypy <- read.csv("../data/schizotypy/share_430n_interview.csv")
```

Data wrangling

```r
# Fill in `NA` with 0
schizotypy[,4:63][is.na(schizotypy[,4:63])] <- 0
```

Implement accuracy function

```r
# R-squared and RMSE
continuous_accuracy <- function(prediction, observed)
{

  # Compute square error
  square_error <- (prediction - observed)^2

  # Return metrics
  return(
      c(
        R2 = 1 - (
          sum(square_error, na.rm = TRUE) /
          sum((observed - mean(observed, na.rm = TRUE))^2, na.rm = TRUE)
        ),
        RMSE = sqrt(mean(square_error, na.rm = TRUE))
      )
```

```
    )

}
```

Perform linear regression

```
# Estimate model
linear <- lm(gas ~ ., data = schizotypy[,4:64])

# Get predictions
linear_predictions <- predict(linear, newdata = schizotypy)

# Compute metrics
continuous_accuracy(linear_predictions, schizotypy$gas)
```

```
      R2     RMSE
0.321514 8.337094
```

Perform random forest cross-validation

```
# Random forest cross-validation for parameters
store_caret <- train(
  x = schizotypy[,4:63], y = schizotypy$gas,
  method = "ranger", metric = "RMSE",
  trControl = trainControl(method = "cv", number = 5),
  tuneGrid = expand.grid(
    mtry = seq(1, 60, 1), # 1:ncol(data)
    min.node.size = seq(1, 10, 1), # 1-10 is usually good
    splitrule = "variance" # classification
  ),
  num.trees = 500, # keep at 500 for the initial search
  importance = "impurity" # set up for later
); store_caret
```

```
Tuning parameter'splitrule' was held constant at a value of variance
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were mtry = 2, splitrule = variance
and min.node.size = 7.
```

Perform random forest cross-validation for trees

```
# With the `mtry` and `min.node.size` parameters,
# search over `num.trees`
trees <- c(10, 50, 100, 250, 500, 1000, 1500)

# Store results
results <- vector("list", length(trees))

# Perform cross-validation (will be much faster than before)
for(i in seq_along(trees)){

  # Perform
  results[[i]] <- train(
    x = schizotypy[,4:63], y = schizotypy$gas,
    method = "ranger", metric = "RMSE",
    trControl = trainControl(method = "cv", number = 5),
    tuneGrid = data.frame(
      mtry = 2, splitrule = "variance",
      min.node.size = 7
    ),
    num.trees = trees[i], # search over trees
    importance = "impurity" # set up for later
  )$results

}

# Combine results
combined <- do.call(rbind.data.frame, results)
combined$num.trees <- trees
combined
```

```
  mtry splitrule min.node.size     RMSE  Rsquared      MAE    RMSESD RsquaredSD
1    2  variance             7 9.489385 0.1336262 7.510283 0.8467263 0.08916880
2    2  variance             7 9.111175 0.1881439 7.234783 1.2552387 0.03721982
3    2  variance             7 9.301875 0.1646563 7.345548 1.0882134 0.08001602
4    2  variance             7 9.250018 0.1738579 7.315144 1.1173254 0.08331345
5    2  variance             7 9.210019 0.1760011 7.279936 1.2988537 0.05978634
6    2  variance             7 9.238249 0.1758556 7.252813 0.7594103 0.05894616
7    2  variance             7 9.211173 0.1782929 7.244450 0.7353661 0.05142699
      MAESD num.trees
1 0.4005441        10
2 0.5111963        50
3 0.2131521       100
4 0.3424297       250
5 0.5334170       500
6 0.1808074      1000
7 0.2931122      1500
```

```
# 50 trees has the best accuracy/kappa

# Get final model with {RandomForest}
ranger_model <- ranger(
  x = schizotypy[,4:63], y = schizotypy$gas,
  mtry = 2, splitrule = "variance",
  min.node.size = 7, num.trees = 50,
  importance = "impurity",
  seed = 42
)

# Get predictions
ranger_predictions <- predict(ranger_model, data = schizotypy)$predictions

# Compute metrics
continuous_accuracy(ranger_predictions, schizotypy$gas)
```

```
       R2      RMSE
0.4654335 7.4002282
```

Linear model significance

```r
# Summarize
summary(linear)
```

```

Call:
lm(formula = gas ~ ., data = schizotypy[, 4:64])

Residuals:
    Min      1Q  Median      3Q     Max 
-46.654  -4.820   0.804   5.993  19.361 

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)    
(Intercept) 78.17417    0.97698  80.016 < 0.0000000000000002 ***
PY01        -2.97941    1.59182  -1.872              0.06204 .  
PY02         0.80704    1.31331   0.615              0.53926    
PY03         2.12295    1.46214   1.452              0.14737    
PY04         1.15336    1.24198   0.929              0.35368    
PY05         1.48985    1.42762   1.044              0.29736    
PY06        -1.67607    1.10646  -1.515              0.13068    
PY07        -1.00276    2.00730  -0.500              0.61768    
PY08         1.72905    1.57694   1.096              0.27359    
PY09         1.21310    1.50556   0.806              0.42091    
PY10        -1.23567    1.98095  -0.624              0.53316    
PY11         1.18828    1.11934   1.062              0.28911    
PY12        -1.06037    1.78475  -0.594              0.55279    
PY13        -0.20081    1.83799  -0.109              0.91306    
PY14        -2.65658    1.52448  -1.743              0.08223 .  
PY15         0.07175    1.24941   0.057              0.95423    
PB01        -0.12802    1.51819  -0.084              0.93285    
PB02         2.57177    1.59535   1.612              0.10781    
PB03         1.89917    1.71690   1.106              0.26938    
PB04        -0.14259    1.96918  -0.072              0.94231    
PB05         2.90596    1.75292   1.658              0.09821 .  
PB06        -1.19992    1.67001  -0.719              0.47290    
PB07        -4.97429    1.82219  -2.730              0.00664 ** 
PB08        -0.67122    1.78668  -0.376              0.70737    
PB09        -2.14888    1.89146  -1.136              0.25665    
PB10        -0.48230    1.73574  -0.278              0.78127    
PB11        -0.09177    2.06414  -0.044              0.96456    
PB12        -2.22896    1.84491  -1.208              0.22776    
PB13        -0.57522    1.42738  -0.403              0.68719    
PB14        -1.22411    1.91245  -0.640              0.52252    
PB15        -1.96974    2.00160  -0.984              0.32572    
MI01         1.54959    1.09966   1.409              0.15963    
MI02         1.04904    1.21192   0.866              0.38727    
MI03        -0.48617    1.16233  -0.418              0.67599    
MI04         1.10412    1.36390   0.810              0.41873    
MI05         0.05553    1.12086   0.050              0.96051    
MI06        -2.17248    1.26730  -1.714              0.08732 .  
MI07        -1.42141    1.18681  -1.198              0.23181    
MI08         0.32120    1.14486   0.281              0.77921    
```

```
MI09        -0.88836    1.27894  -0.695              0.48774
MI10        -1.30769    1.11183  -1.176              0.24029
MI11         1.03636    1.15490   0.897              0.37011
MI12        -0.13277    1.52663  -0.087              0.93074
MI13        -0.56922    1.30318  -0.437              0.66252
MI14        -0.57595    1.19729  -0.481              0.63077
MI15        -2.40929    1.82346  -1.321              0.18723
SA01         0.66409    1.55297   0.428              0.66917
SA02        -2.59132    1.48950  -1.740              0.08274 .
SA03        -1.51302    1.44981  -1.044              0.29735
SA04        -1.13573    1.91499  -0.593              0.55349
SA05        -1.09718    1.49654  -0.733              0.46393
SA06        -2.06251    1.35975  -1.517              0.13017
SA07        -1.85558    1.37521  -1.349              0.17807
SA08        -3.96097    1.35203  -2.930              0.00360 **
SA09         1.02359    1.73769   0.589              0.55619
SA10        -0.36921    1.29016  -0.286              0.77491
SA11        -1.59763    1.33111  -1.200              0.23082
SA12         1.14421    1.54555   0.740              0.45957
SA13        -1.14477    1.34565  -0.851              0.39548
SA14         0.92547    1.48099   0.625              0.53242
SA15         1.74924    1.66762   1.049              0.29489
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9 on 369 degrees of freedom
Multiple R-squared:  0.3215,    Adjusted R-squared:  0.2112
F-statistic: 2.914 on 60 and 369 DF,  p-value: 0.0000000003457
```

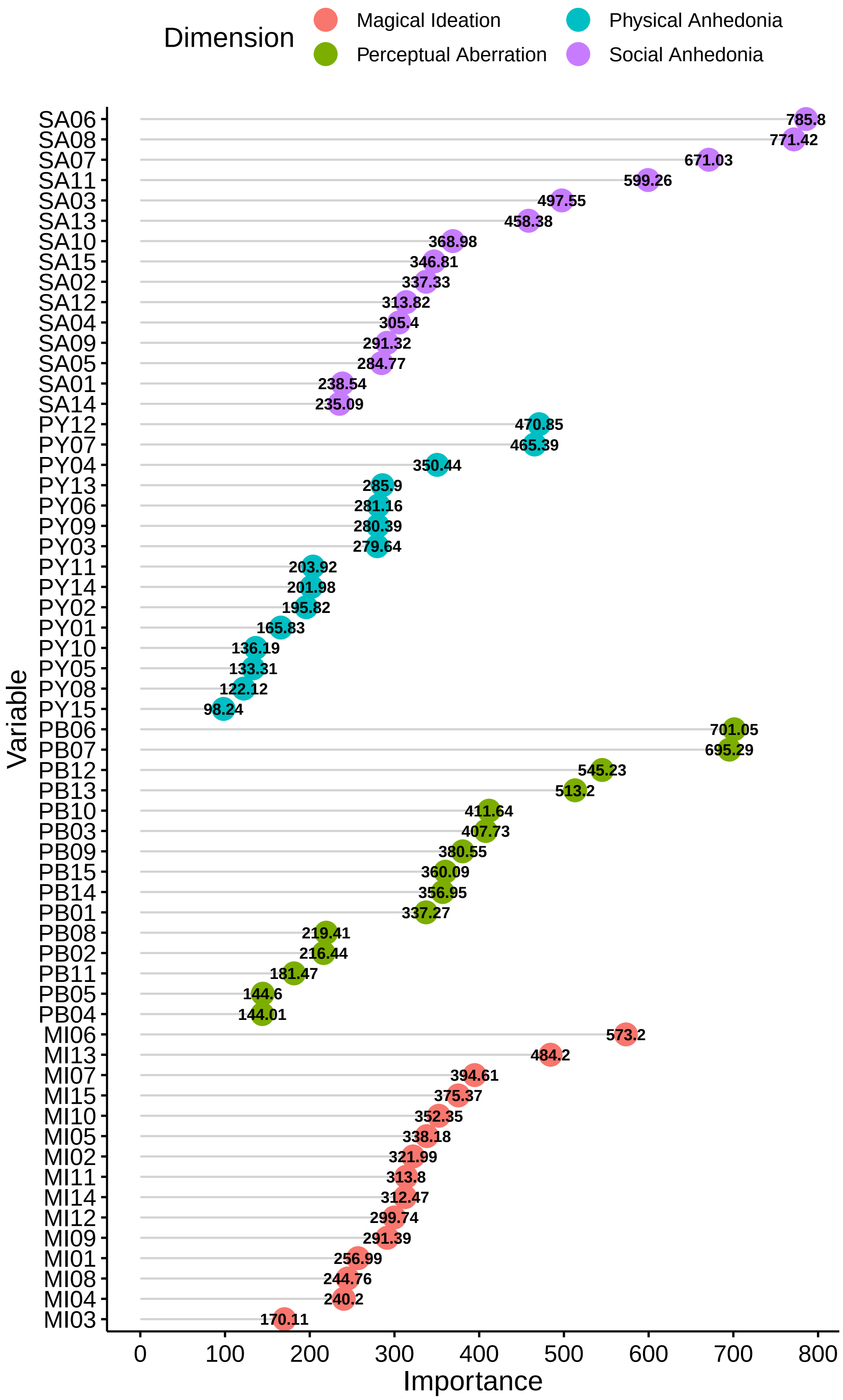
```r
# Significant variables
cat(paste0("Significant: PB07, SA08"))
```

```
Significant: PB07, SA08
```

Relative importance

```r
# Check out importance
ranger_imp <- importance(ranger_model)

# Visualize importance
ggdotchart(
  data = data.frame(
    Importance = round(ranger_imp, 2),
    Variable = names(ranger_imp),
    Dimension = rep(
      c(
        "Physical Anhedonia", "Perceptual Aberration",
        "Magical Ideation", "Social Anhedonia"
      ), each = 15
    )
  ),
  x = "Variable", y = "Importance", color = "Dimension",
  dot.size = 5, add = "segments", label = "Importance",
  group = "Dimension", # for within-dimension comparison
  font.label = list(size = 8, vjust = 0.5, color = "black", face = "bold")
) +
  scale_y_continuous(n.breaks = 8) +
  guides(color = guide_legend(title.position = "left", nrow = 2)) +
  theme(
    legend.position = "top",
    legend.title = element_text(size = 14, hjust = 0.5),
    legend.text = element_text(size = 10),
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12),
    axis.text.x = element_text(angle = 0, hjust = 0.5)
  ) +
  coord_flip()
```


Dimension
● Magical Ideation
● Perceptual Aberration
● Physical Anhedonia
● Social Anhedonia
Variable
Importance
0
100
200
300
400
500
600
700
800
SA06 785.8
SA08 771.42
SA07 671.03
SA11 599.26
SA03 497.55
SA13 458.38
SA10 368.98
SA15 346.81
SA02 337.33
SA12 313.82
SA04 305.4
SA09 291.32
SA05 284.77
SA01 238.54
SA14 235.09
PY12 470.85
PY07 465.39
PY04 350.44
PY13 285.9
PY06 281.16
PY09 280.39
PY03 279.64
PY11 203.92
PY14 201.98
PY02 195.82
PY01 165.83
PY10 136.19
PY05 133.31
PY08 122.12
PY15 98.24
PB06 701.05
PB07 695.29
PB12 545.23
PB13 513.2
PB10 411.64
PB03 407.73
PB09 380.55
PB15 360.09
PB14 356.95
PB01 337.27
PB08 219.41
PB02 216.44
PB11 181.47
PB05 144.6
PB04 144.01
MI06 573.2
MI13 484.2
MI07 394.61
MI15 375.37
MI10 352.35
MI05 338.18
MI02 321.99
MI11 313.8
MI14 312.47
MI12 299.74
MI09 291.39
MI01 256.99
MI08 244.76
MI04 240.2
MI03 170.11

Model comparison

Linear model

- Significant: PB07, SA08
- $R^2 = 0.322$; $RMSE = 8.337$

Random forest

- Importance (> 600): SA06, SA07, SA08, PB06, PB07
- $R^2 = 0.465$; $RMSE = 7.400$

  I prefer the random forest model here because it has nearly 15% greater variance explained (0.143) over the linear model. The relative importance has a few more variables worth noting meaning the relationship between our predictors and outcome is likely to be nonlinear.

Random Forest Explainer

Vignette on how to use {randomForestExplainer}

```
# Load {randomForestExplainer}
library(randomForestExplainer)

# Explain forest (produces an HTML file)
explain_forest(ranger_model, data = schizotypy)
```