permute_rf_strobl_xor_mtry

Kelvyn Bladen

2024-05-17

```
s <- Sys.time()
library(ggplot2)
library(GGally)
library(ggeasy)
library(randomForest)
library(dplyr)
library(randomForestVIP)
library(tidyr)
```

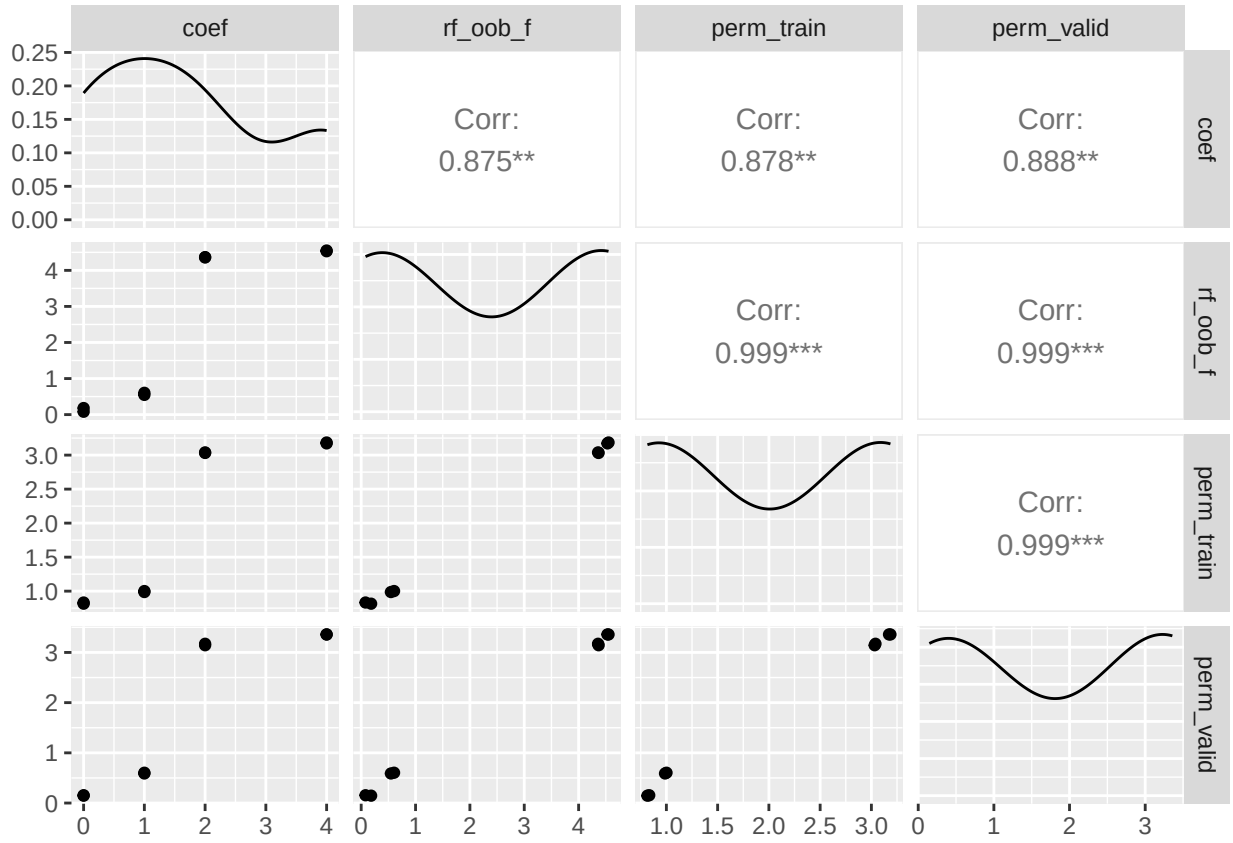
```
rsq = vector(length = 8)

rf_oob_t <- mat.or.vec(8, 8)
rf_oob_f <- mat.or.vec(8, 8)
# rf_pdp <- mat.or.vec(8, 8)

perm_train <- mat.or.vec(8, 8)
drop_train <- mat.or.vec(8, 8)

perm_valid <- mat.or.vec(8, 8)
drop_valid <- mat.or.vec(8, 8)

mrep <- 20
n_size = 1000
set.seed(123)

for (j in seq_len(mrep)) {
    sig <- diag(1, 8, 8)

    for (ii in 1:4) {
      for (jj in 1:4) {
        sig[ii, jj] <- ifelse(ii == jj, 1, 0.95)
      }
    }

    strobl <- MASS::mvrnorm(n_size, mu = rep(0, 8), Sigma = sig)

    y <- 4 * strobl[, 1]*strobl[, 2] + 2 * strobl[, 3]*strobl[, 4] +
      strobl[, 5]*strobl[, 6] + rnorm(n_size, mean = 0, sd = .1)
    strobl <- data.frame(cbind(strobl, y))

    dfv <- MASS::mvrnorm(n_size, mu = rep(0, 8), Sigma = sig)
    y <- 4 * dfv[, 1]*dfv[, 2] + 2 * dfv[, 3]*dfv[, 4] +
```

```
    dfv[, 5]*dfv[, 6] + rnorm(n_size, mean = 0, sd = .1)
  dfv <- data.frame(cbind(dfv, y))

for (k  in seq_len(8)) {
  r <- randomForest(y ~ ., data = strobl, mtry = k,
                    importance = T)

  impt <- sqrt(as.data.frame(pmax(randomForest::importance(r, scale = T), 0)))
  impt <- impt$`%IncMSE`[1:8]

  impf <- sqrt(as.data.frame(pmax(randomForest::importance(r, scale = F), 0)))
  impf <- impf$`%IncMSE`[1:8]

  # vimp = pdp_compare(r, var_vec = 1:8, trellis = F)
  # impp = vimp$imp[c(1, 4)] %>% arrange(var) %>% pull(sd)

  # vimp = vip::vi_firm(r, train = strobl)
  # impp <- vimp$Importance[1:8]

  p <- predict(r, strobl)
  m = mean((p-strobl$y)^2)

  rq = r$rsq[500]

  vp <- predict(r, dfv)
  mv = mean((vp-dfv$y)^2)

  perm_impr <- vector(length = 8)
  perm_impv <- vector(length = 8)
  drop_impr <- vector(length = 8)
  drop_impv <- vector(length = 8)

  for (i in seq_len(8)) {
    df_new <- strobl
    df_new[i] <- df_new[sample(1:n_size), i]

    p <- predict(r, df_new)
    new_m = mean((p-strobl$y)^2)
    perm_impr[i] <- new_m - m

    ##########################################

    v_new <- dfv
    v_new[i] <- v_new[sample(1:n_size), i]

    vp <- predict(r, v_new)
    new_vm = mean((vp-dfv$y)^2)
    perm_impv[i] <- new_vm - mv

    ##########################################

    df_new <- strobl
    df_new[, i] <- 0
```

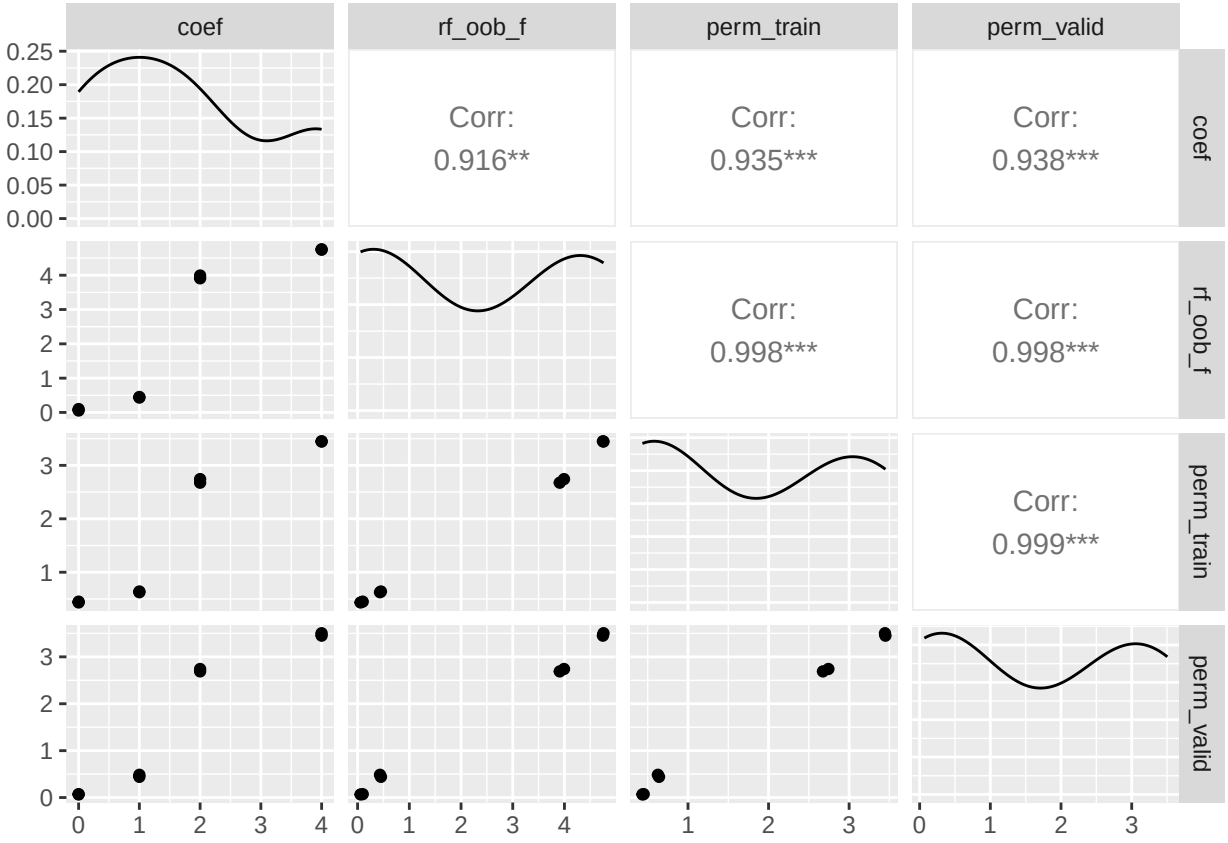
```
      p <- predict(r, df_new)
      new_m = mean((p-strobl$y)^2)
      drop_impr[i] <- new_m - m

      ################################################

      v_new <- dfv
      v_new[, i] <- 0

      vp <- predict(r, v_new)
      new_vm = mean((vp-dfv$y)^2)
      drop_impv[i] <- new_vm - mv
    }

    rf_oob_t[,k] <- rf_oob_t[,k] + impt / mrep
    rf_oob_f[,k] <- rf_oob_f[,k] + impf / mrep

    # rf_pdp[,k] <- rf_pdp[,k] + impp / mrep

    rsq[k] <- rsq[k] + rq / mrep

    simpr <- sqrt(pmax(perm_impr, 0))
    perm_train[,k] <- perm_train[,k] + simpr / mrep

    simpv <- sqrt(pmax(perm_impv, 0))
    perm_valid[,k] <- perm_valid[,k] + simpv / mrep

    dsimpr <- sqrt(pmax(drop_impr, 0))
    drop_train[,k] <- drop_train[,k] + dsimpr / mrep

    dsimpv <- sqrt(pmax(drop_impv, 0))
    drop_valid[,k] <- drop_valid[,k] + dsimpv / mrep
  }

}
```

```
for (i in seq_len(8)){
  sdf <- data.frame(coef = c(4,4,2,2,1,1,0,0),
                    rf_oob_t = rf_oob_t[,i],
                    rf_oob_f = rf_oob_f[,i],
                    # rf_pdp = rf_pdp[,i],
                    perm_train = perm_train[,i],
                    drop_train = drop_train[,i],
                    perm_valid = perm_valid[,i],
                    drop_valid = drop_valid[,i])

  sdf <- sdf %>% select(coef, rf_oob_f, #rf_pdp,
                        perm_train, perm_valid)
  print(sdf)
  print(ggpairs(sdf))
}
```

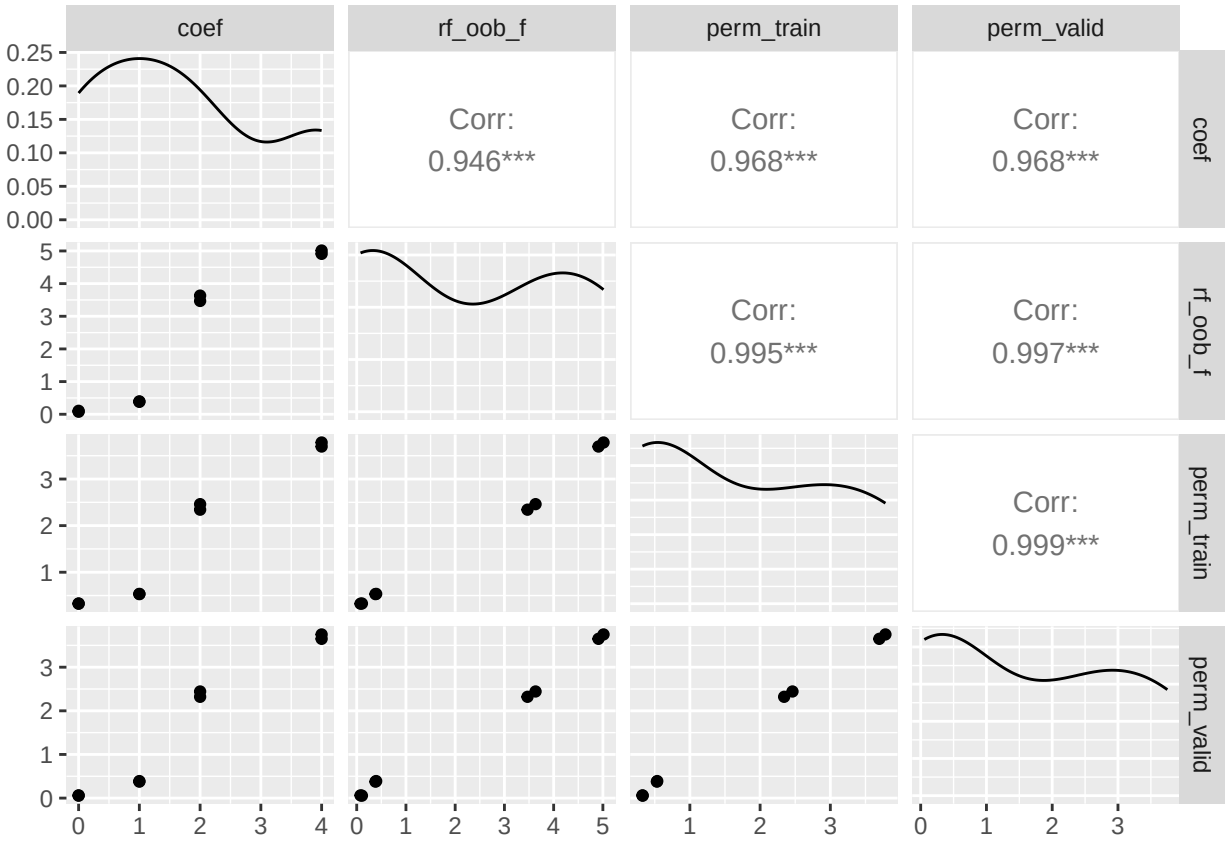
```
##   coef   rf_oob_f perm_train perm_valid
## 1    4 4.55023521  3.1856827  3.3579206
## 2    4 4.53027369  3.1720421  3.3586543
## 3    2 4.35784725  3.0397729  3.1726274
## 4    2 4.36764848  3.0319075  3.1435726
## 5    1 0.54661396  0.9869110  0.5888829
## 6    1 0.60464068  1.0012861  0.6039617
## 7    0 0.07965812  0.8328667  0.1557269
## 8    0 0.18264961  0.8152812  0.1471168
```



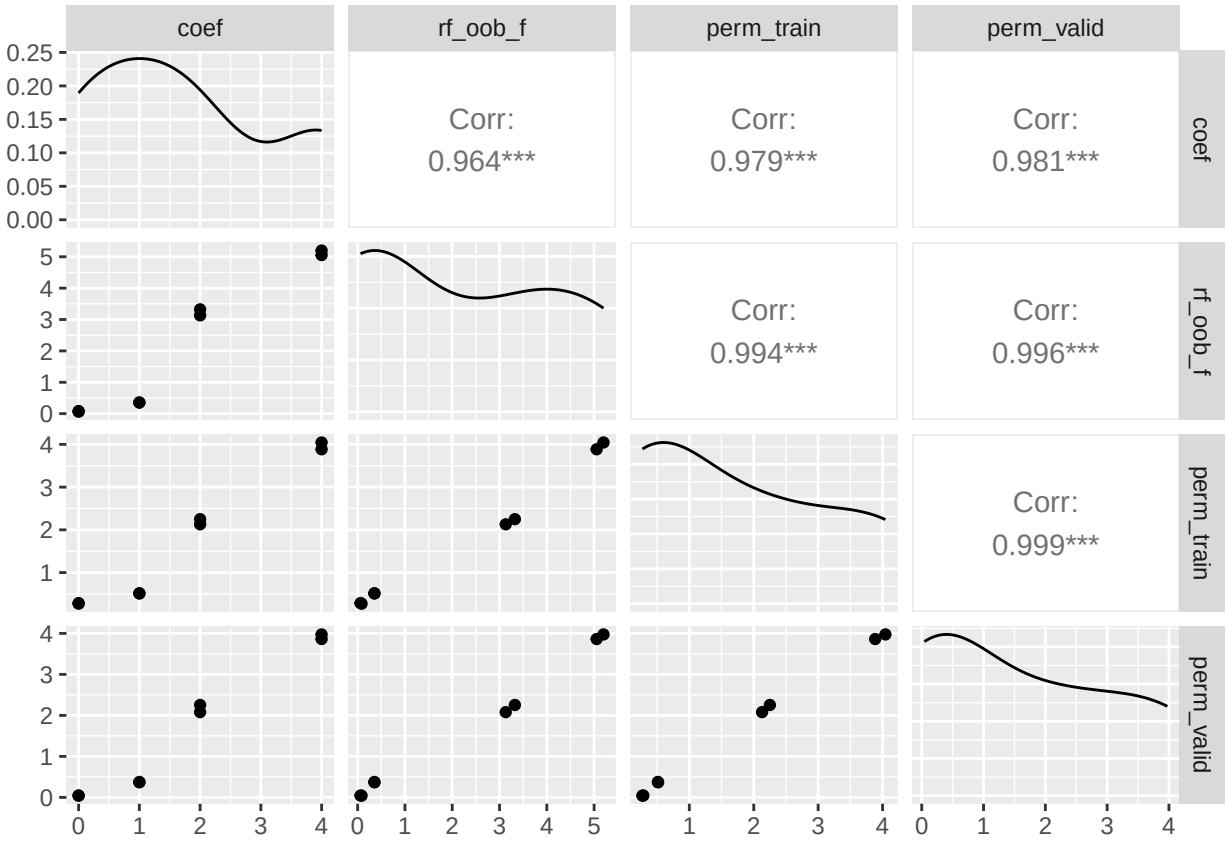
```
##   coef   rf_oob_f perm_train perm_valid
## 1    4 4.74393071  3.4510128 3.45222762
## 2    4 4.75571919  3.4451625 3.50432158
## 3    2 3.99144675  2.7426322 2.74144221
## 4    2 3.90929388  2.6752925 2.69019706
## 5    1 0.45146335  0.6402865 0.44172669
## 6    1 0.43302786  0.6262605 0.48565418
## 7    0 0.05638079  0.4343187 0.06639882
## 8    0 0.09886792  0.4498277 0.07233167
```



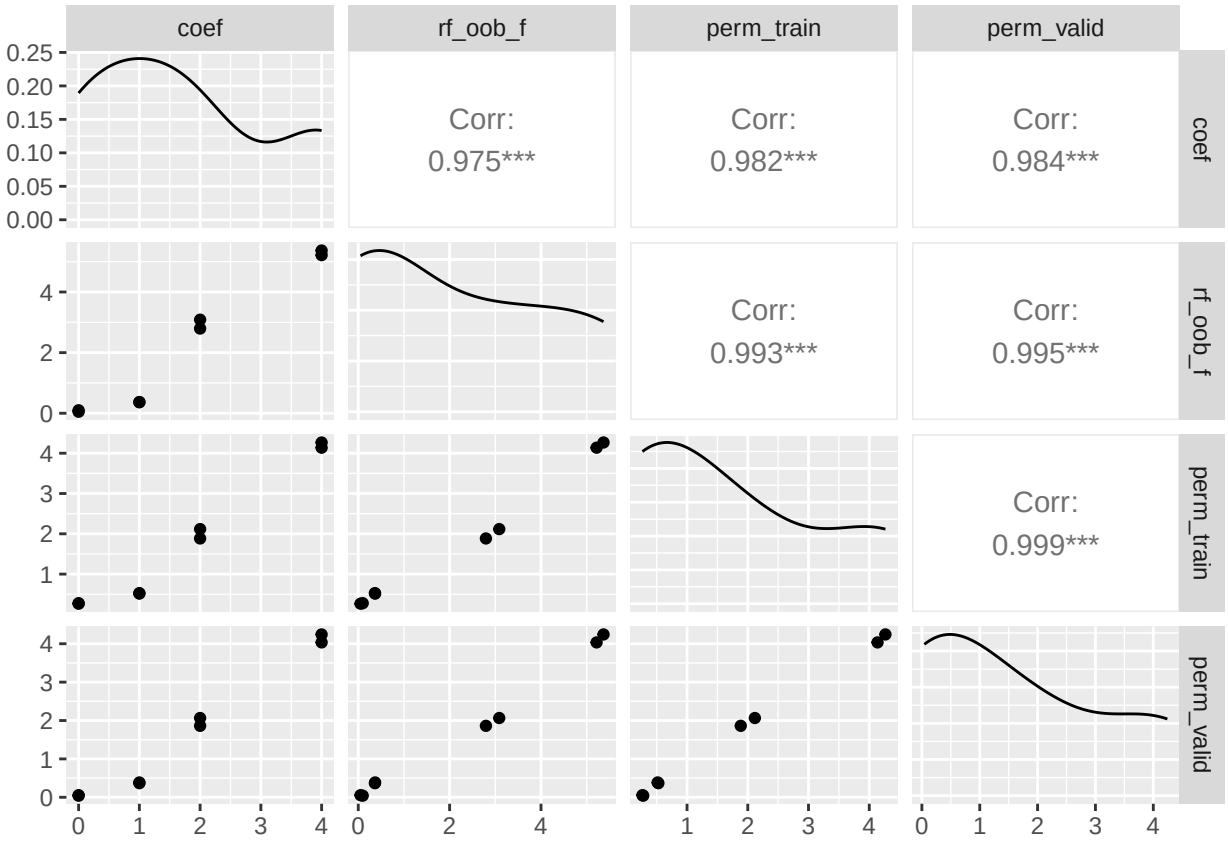

```
##   coef   rf_oob_f perm_train perm_valid
## 1    4 4.91104986  3.6969829 3.64969426
## 2    4 5.01309897  3.7842390 3.75056498
## 3    2 3.63280087  2.4609864 2.44358702
## 4    2 3.46916235  2.3421430 2.32090904
## 5    1 0.39506086  0.5352146 0.38993900
## 6    1 0.38088652  0.5303850 0.37693187
## 7    0 0.07732107  0.3246554 0.06255278
## 8    0 0.10552855  0.3289655 0.05575746
```



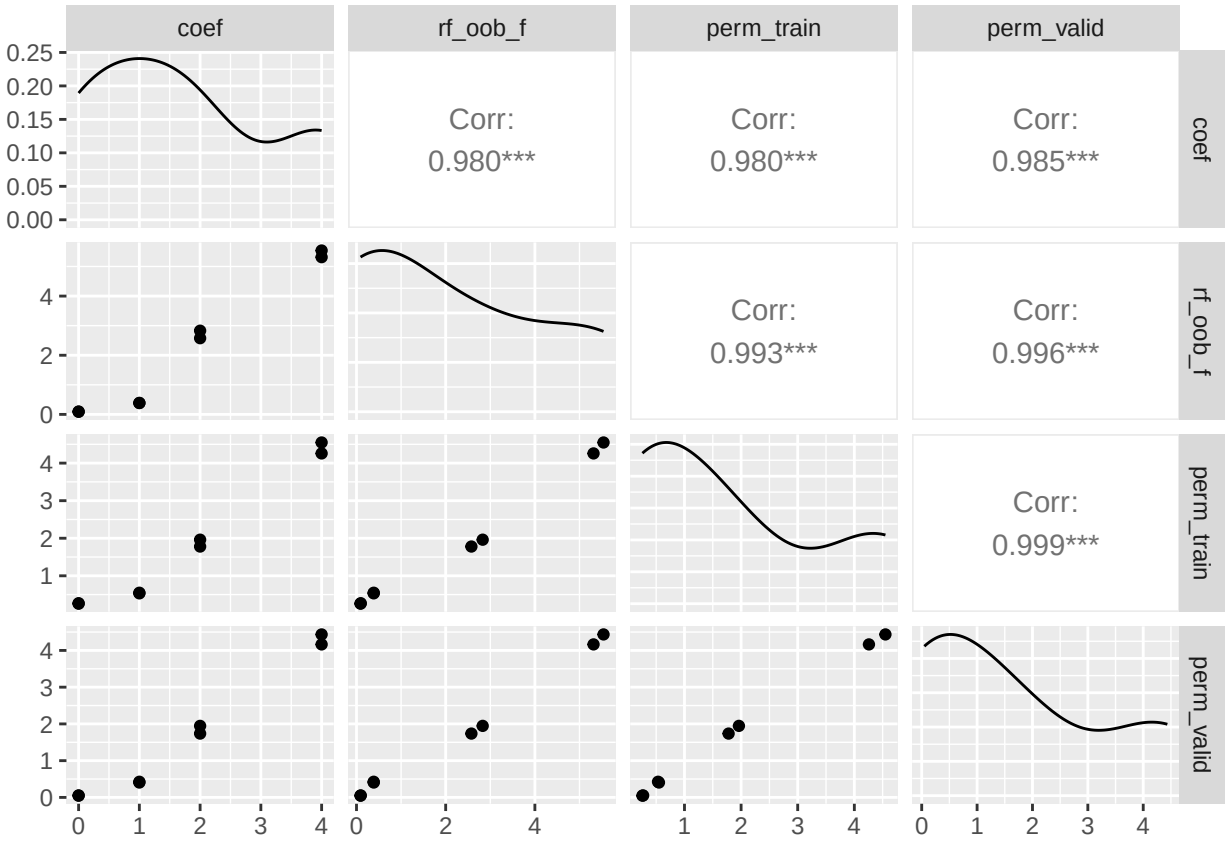

```
##   coef   rf_oob_f perm_train perm_valid
## 1    4 5.05382741  3.8846589 3.86245704
## 2    4 5.19942261  4.0442194 3.97414343
## 3    2 3.32469532  2.2503990 2.25367588
## 4    2 3.13165860  2.1294429 2.07891646
## 5    1 0.35695834  0.5151031 0.36742044
## 6    1 0.35331820  0.5117834 0.37611905
## 7    0 0.08299648  0.2718829 0.04216365
## 8    0 0.06095418  0.2858855 0.04543980
```



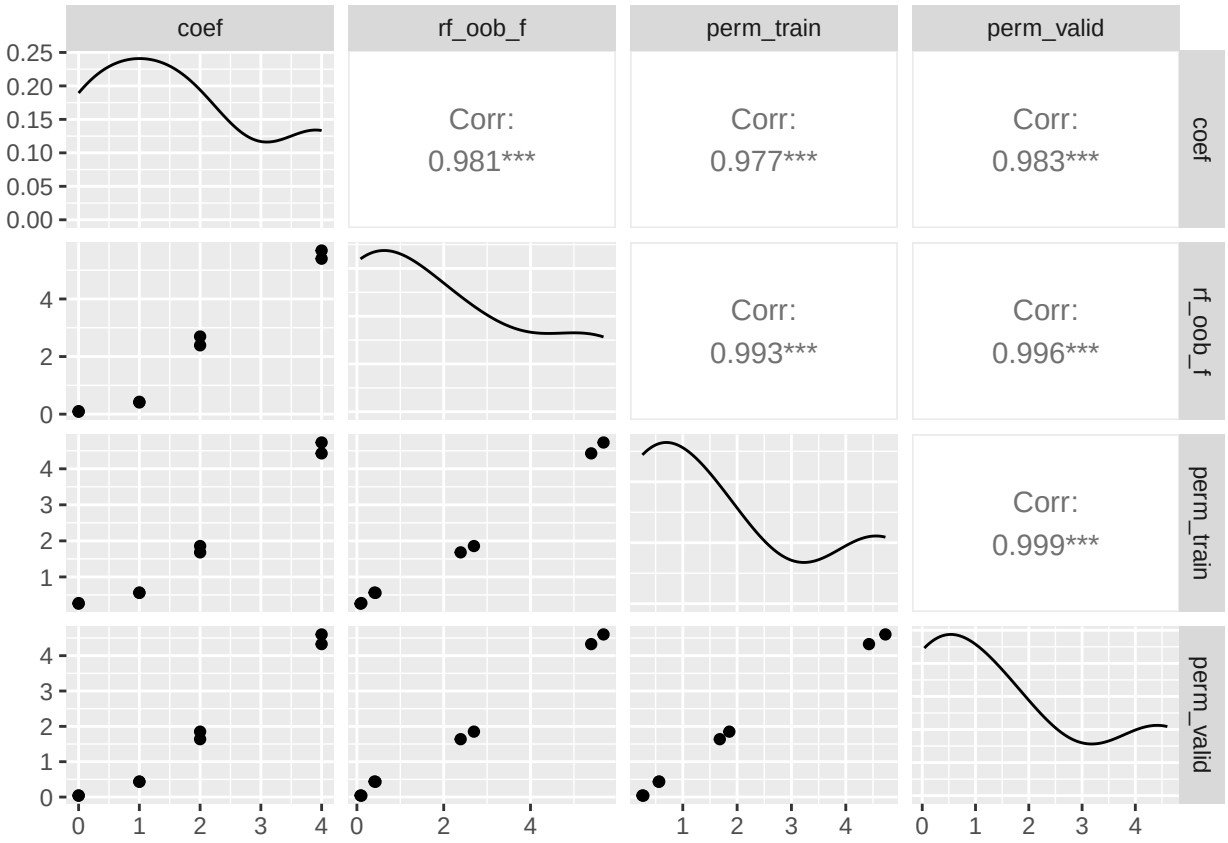

```
##   coef   rf_oob_f perm_train perm_valid
## 1    4 5.21714967  4.1369358 4.03486352
## 2    4 5.36849879  4.2664109 4.24272738
## 3    2 3.08595658  2.1167134 2.06567165
## 4    2 2.79437023  1.8847290 1.86175423
## 5    1 0.36411390  0.5279632 0.36776503
## 6    1 0.36941979  0.5166259 0.38605868
## 7    0 0.09756638  0.2785194 0.04240926
## 8    0 0.04967765  0.2657587 0.05461042
```



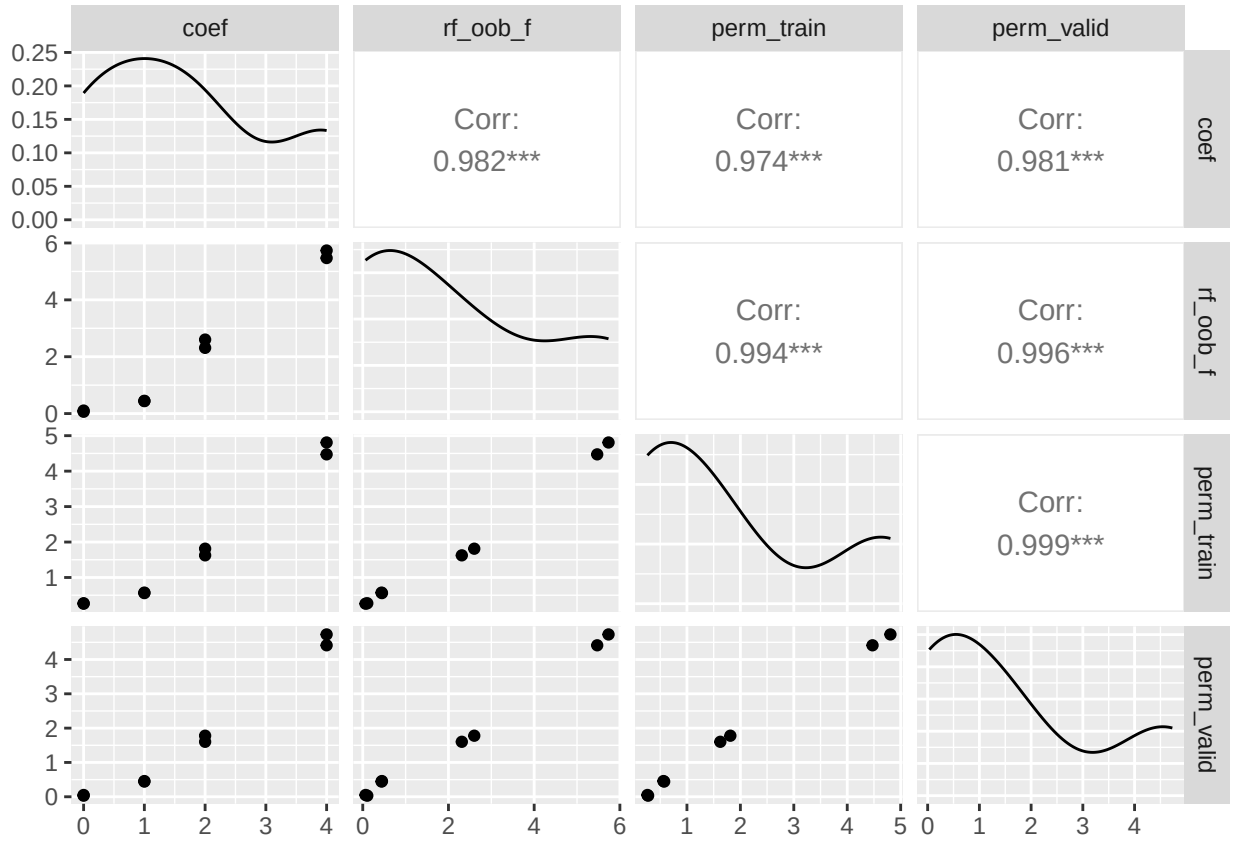

```
##   coef   rf_oob_f perm_train perm_valid
## 1    4 5.31565088  4.2577892 4.16398935
## 2    4 5.53971149  4.5492465 4.43532117
## 3    2 2.82983117  1.9614922 1.94712171
## 4    2 2.57640301  1.7793785 1.73750440
## 5    1 0.38662407  0.5492701 0.40566615
## 6    1 0.38552652  0.5332681 0.42614654
## 7    0 0.09521966  0.2699557 0.05366484
## 8    0 0.09009455  0.2575570 0.04786698
```




```
##   coef   rf_oob_f perm_train perm_valid
## 1    4 5.39848265  4.4275322 4.32717329
## 2    4 5.68235578  4.7298687 4.60076353
## 3    2 2.69652660  1.8568266 1.85108186
## 4    2 2.39054078  1.6813107 1.63746720
## 5    1 0.40606859  0.5658059 0.43946923
## 6    1 0.42895723  0.5570602 0.43205608
## 7    0 0.10268457  0.2745382 0.03937104
## 8    0 0.08335317  0.2572054 0.04436846
```




```
##   coef   rf_oob_f perm_train perm_valid
## 1    4 5.47130018  4.4732624 4.41394297
## 2    4 5.73529556  4.8084934 4.73036273
## 3    2 2.60242442  1.8122576 1.77893935
## 4    2 2.31160261  1.6244814 1.60175336
## 5    1 0.43867794  0.5734284 0.44358671
## 6    1 0.44786901  0.5599756 0.45731382
## 7    0 0.10548432  0.2731130 0.03248328
## 8    0 0.06453655  0.2605357 0.05162494
```

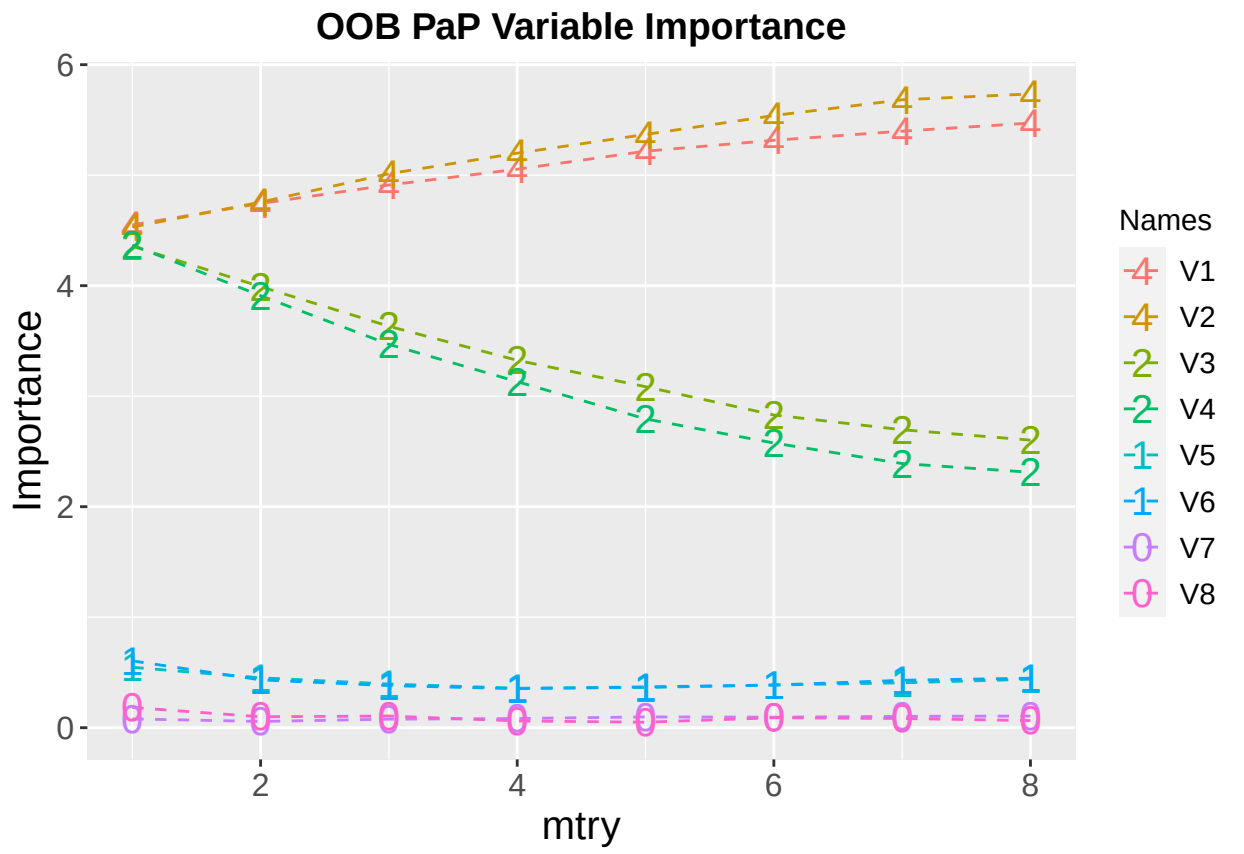



```
rsq
```

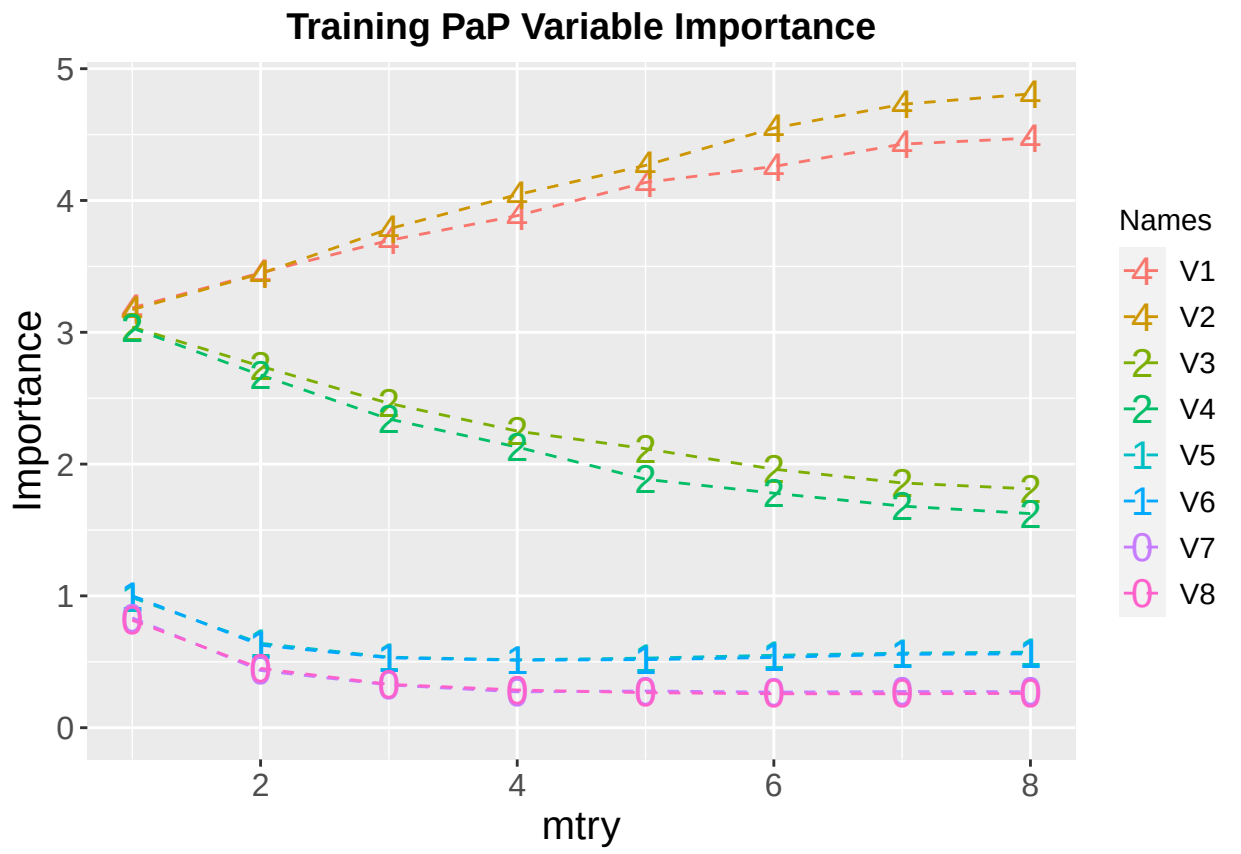
```
## [1] 0.9299820 0.9587742 0.9641311 0.9654866 0.9660085 0.9656591 0.9654013
## [8] 0.9646779
```

```
Names = c("V1", "V2", "V3", "V4", "V5", "V6", "V7", "V8")

mag <- dplyr::case_when(Names %in% c("V1", "V2") ~ 4,
                  Names %in% c("V3", "V4") ~ 2,
                  Names %in% c("V5", "V6") ~ 1,
                  .default = 0)

Names <- factor(Names,
                levels = c("V1", "V2", "V3", "V4",
                           "V5","V6","V7","V8"), ordered = T)
Names <- factor(Names, ordered = F)

rf_oob_f1 = data.frame(rf_oob_f, Names, mag)
# rf_pdp1 = data.frame(rf_pdp, Names, mag)
perm_train1 = data.frame(perm_train, Names, mag)
drop_valid1 = data.frame(drop_valid, Names, mag)
perm_valid1 = data.frame(perm_valid, Names, mag)

colnames(rf_oob_f1)[1:8] <- 1:8
rf_oob_f1 <- rf_oob_f1 %>% pivot_longer(!c(Names,mag), names_to = "mtry",
```

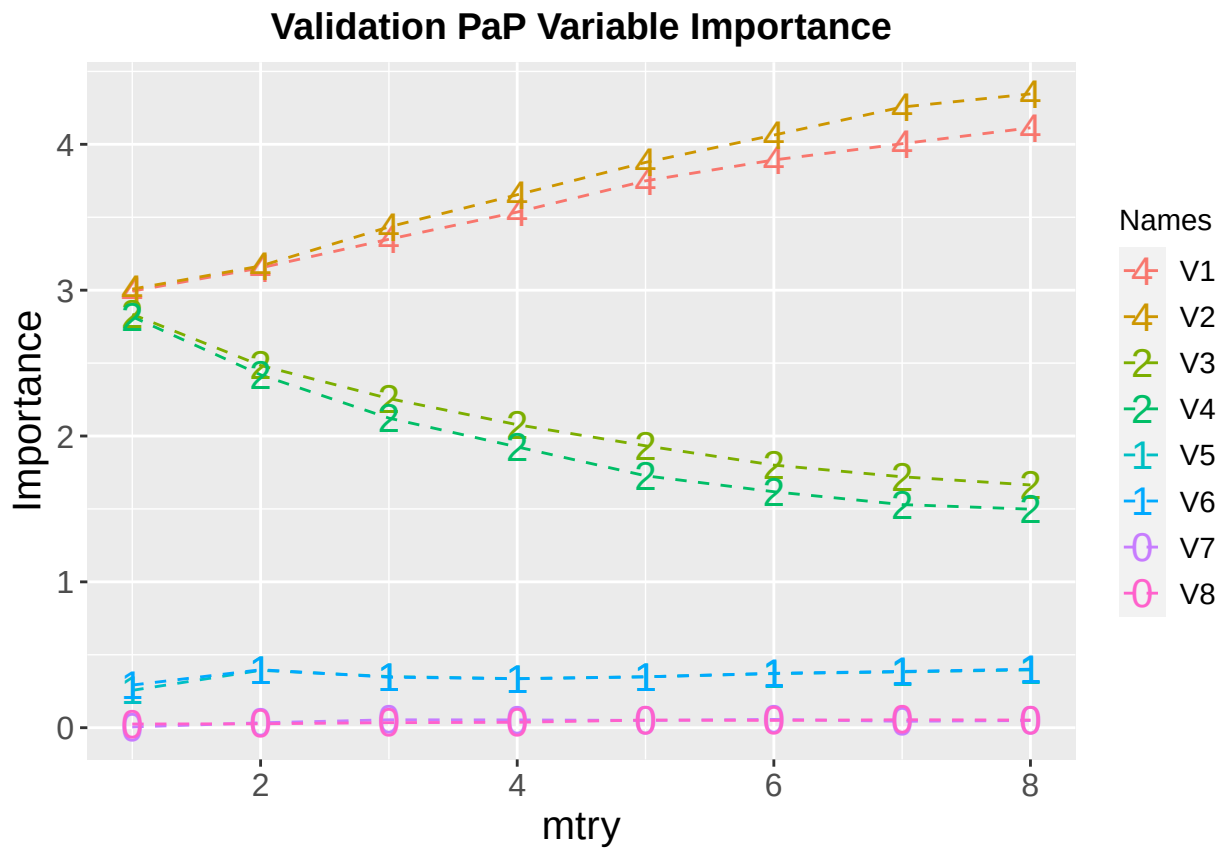
```
                                        values_to = "Imp")
rf_oob_f1$mtry <- as.numeric(rf_oob_f1$mtry)

# colnames(rf_pdp1)[1:8] <- 1:8
# rf_pdp1 <- rf_pdp1 %>% pivot_longer(!c(Names,mag), names_to = "mtry",
#                                     values_to = "Imp")
# rf_pdp1$mtry <- as.numeric(rf_pdp1$mtry)

colnames(perm_train1)[1:8] <- 1:8
perm_train1 <- perm_train1 %>%
  pivot_longer(!c(Names,mag), names_to = "mtry", values_to = "Imp")
perm_train1$mtry <- as.numeric(perm_train1$mtry)

colnames(drop_valid1)[1:8] <- 1:8
drop_valid1 <- drop_valid1 %>%
  pivot_longer(!c(Names,mag), names_to = "mtry", values_to = "Imp")
drop_valid1$mtry <- as.numeric(drop_valid1$mtry)

colnames(perm_valid1)[1:8] <- 1:8
perm_valid1 <- perm_valid1 %>%
  pivot_longer(!c(Names,mag), names_to = "mtry", values_to = "Imp")
perm_valid1$mtry <- as.numeric(perm_valid1$mtry)

ma = max(rf_oob_f1$Imp, perm_train1$Imp, perm_valid1$Imp, drop_valid1$Imp)
# mp = max(rf_pdp1$Imp)

gr <- rf_oob_f1 %>%
  ggplot(aes(x = mtry, y = Imp, color = Names,
             group = Names, linetype = Names,
             shape = Names)) +
  geom_line() +
  scale_x_continuous(limits = c(1,8), breaks = seq(2,8,by=2)) +
  scale_y_continuous(limits = c(0,max(rf_oob_f1$Imp))) +
  #scale_y_continuous(limits = c(0,4), breaks = seq(0,4,by=1)) +
  ggtitle("OOB PaP Variable Importance") +
  geom_point(size = 5) +
  scale_linetype_manual(values = rep(2, each = 8)) +
  scale_shape_manual(values = c(52,52,50,50,49,49,48,48)) +
  scale_size(range = c(6,6)) +
  ylab("Importance") +
  guides(size = "none") +
  theme(axis.text = element_text(size = 12),
        axis.title = element_text(size = 15),
        plot.title = element_text(size = 14, face = "bold")) +
  easy_center_title() + easy_plot_legend_size(size = 11)
gr
```



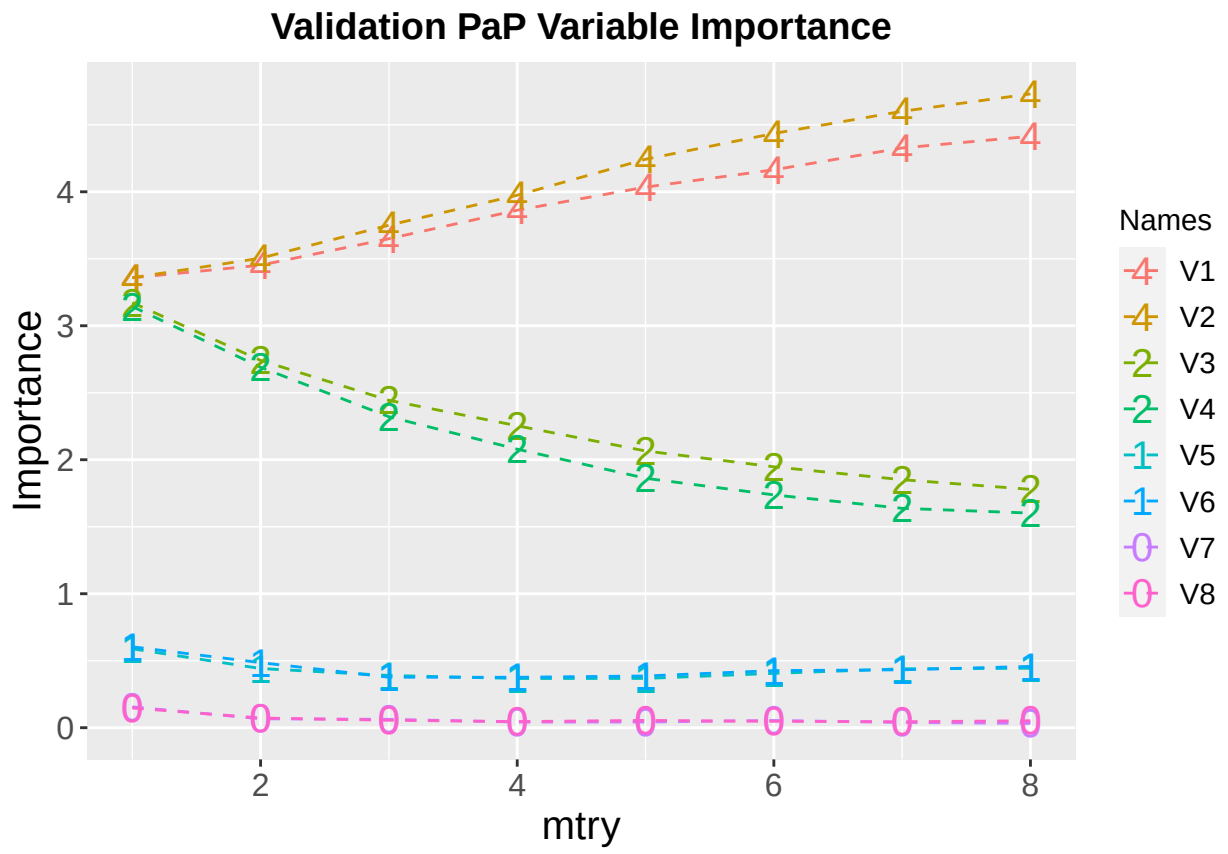

```
# ggsave("xor_oob_zoom.pdf", plot = last_plot(), dpi = 2400,
#        width = 6, height = 6)

gp <- perm_train1 %>%
  ggplot(aes(x = mtry, y = Imp, color = Names,
             group = Names, linetype = Names,
             shape = Names)) +
  geom_line() +
  scale_x_continuous(limits = c(1,8), breaks = seq(2,8,by=2)) +
  scale_y_continuous(limits = c(0,max(perm_train1$Imp))) +
  #scale_y_continuous(limits = c(0,4), breaks = seq(0,4,by=1)) +
  ggtitle("Training PaP Variable Importance") +
  geom_point(size = 5) +
  scale_linetype_manual(values = rep(2, each = 8)) +
  scale_shape_manual(values = c(52,52,50,50,49,49,48,48)) +
  scale_size(range = c(6,6)) +
  ylab("Importance") +
  guides(size = "none") +
  theme(axis.text = element_text(size = 12),
        axis.title = element_text(size = 15),
        plot.title = element_text(size = 14, face = "bold")) +
  easy_center_title() + easy_plot_legend_size(size = 11)
gp
```




```
# ggsave("xor_train_zoom.pdf", plot = last_plot(), dpi = 2400,
#        width = 6, height = 6)

gd <- drop_valid1 %>%
  ggplot(aes(x = mtry, y = Imp, color = Names,
             group = Names, linetype = Names,
             shape = Names)) +
  geom_line() +
  scale_x_continuous(limits = c(1,8), breaks = seq(2,8,by=2)) +
  scale_y_continuous(limits = c(0,max(drop_valid1$Imp))) +
  #scale_y_continuous(limits = c(0,4), breaks = seq(0,4,by=1)) +
  ggtitle("Validation PaP Variable Importance") +
  geom_point(size = 5) +
  scale_linetype_manual(values = rep(2, each = 8)) +
  scale_shape_manual(values = c(52,52,50,50,49,49,48,48)) +
  scale_size(range = c(6,6)) +
  ylab("Importance") +
  guides(size = "none") +
  theme(axis.text = element_text(size = 12),
        axis.title = element_text(size = 15),
        plot.title = element_text(size = 14, face = "bold")) +
  easy_center_title() + easy_plot_legend_size(size = 11)
gd
```

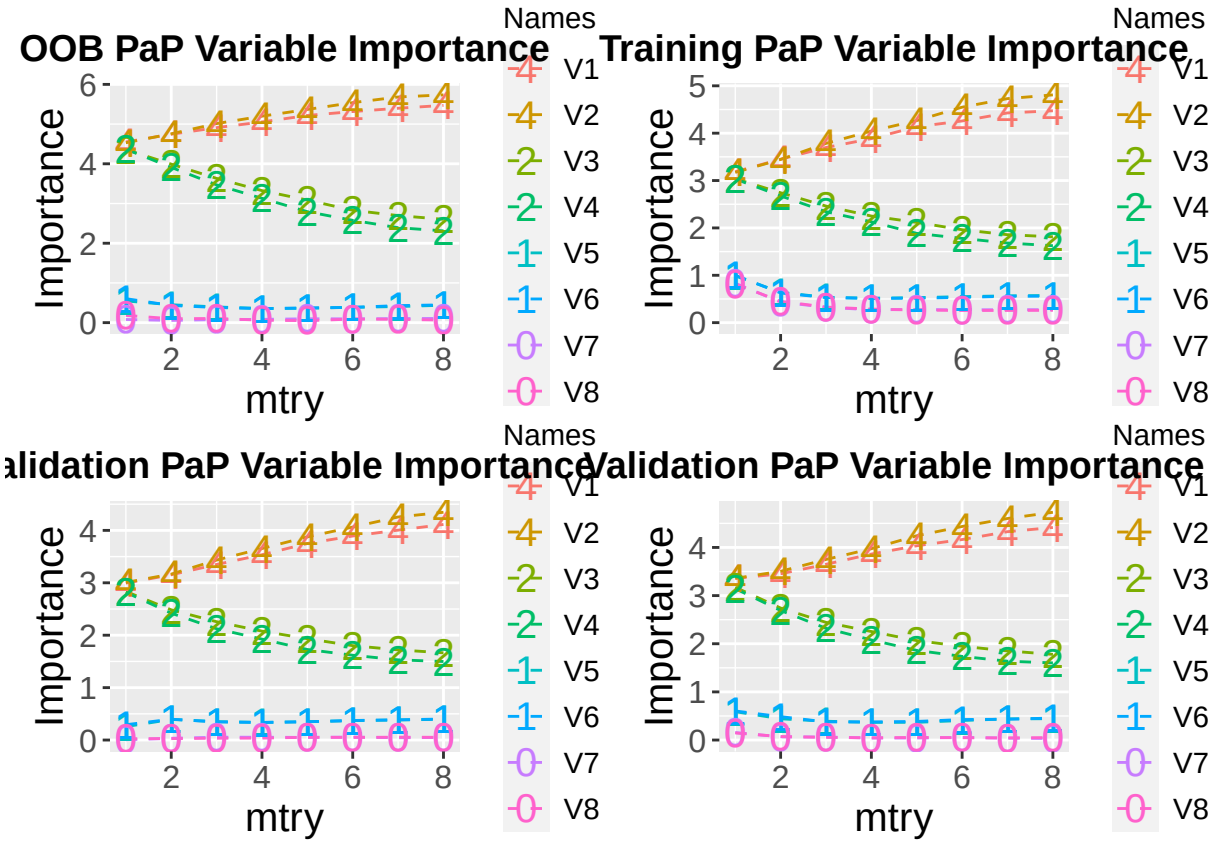



```
gv <- perm_valid1 %>%
  ggplot(aes(x = mtry, y = Imp, color = Names,
             group = Names, linetype = Names,
             shape = Names)) +
  geom_line() +
  scale_x_continuous(limits = c(1,8), breaks = seq(2,8,by=2)) +
  scale_y_continuous(limits = c(0,max(perm_valid1$Imp))) +
  #scale_y_continuous(limits = c(0,4), breaks = seq(0,4,by=1)) +
  ggtitle("Validation PaP Variable Importance") +
  geom_point(size = 5) +
  scale_linetype_manual(values = rep(2, each = 8)) +
  scale_shape_manual(values = c(52,52,50,50,49,49,48,48)) +
  scale_size(range = c(6,6)) +
  ylab("Importance") +
  guides(size = "none") +
  theme(axis.text = element_text(size = 12),
        axis.title = element_text(size = 15),
        plot.title = element_text(size = 14, face = "bold")) +
  easy_center_title() + easy_plot_legend_size(size = 11)
gv
```



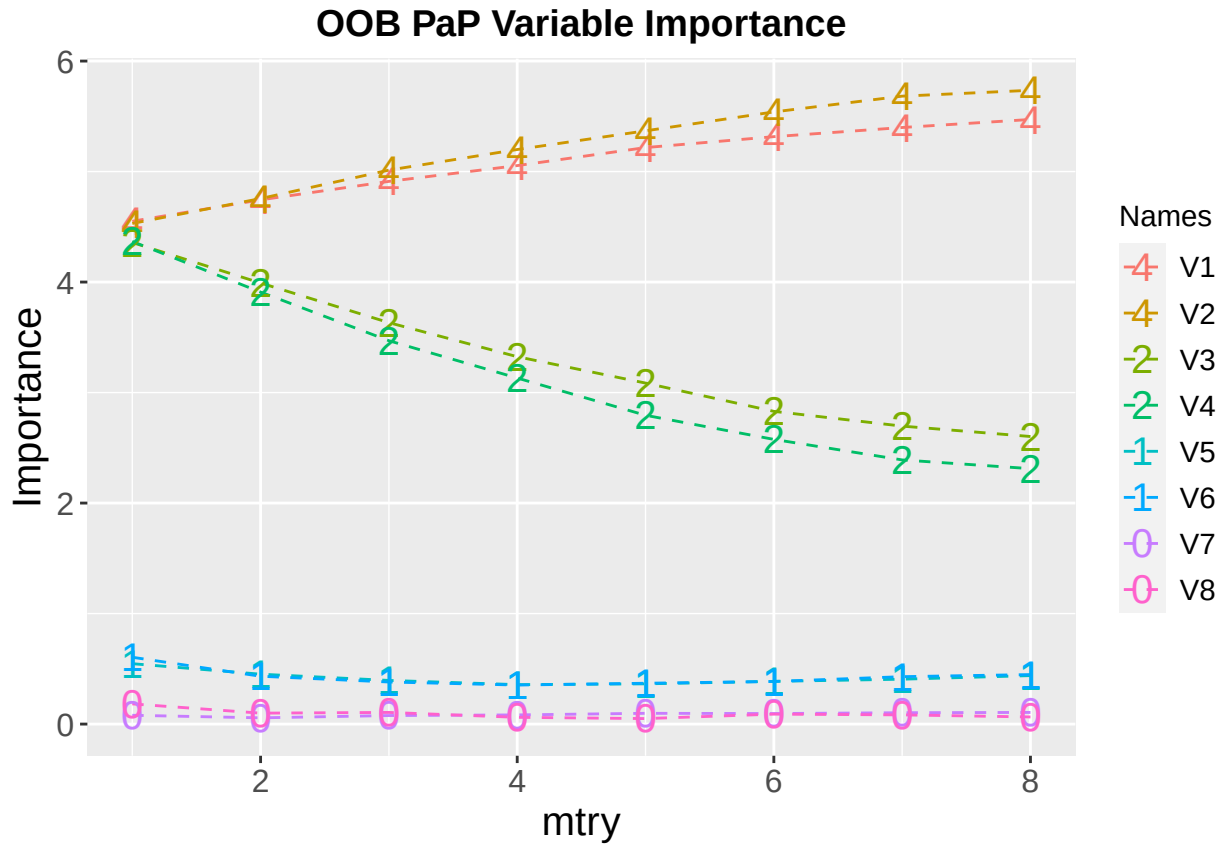

```
# ggsave("xor_val_zoom.pdf", plot = last_plot(), dpi = 2400,
#        width = 6, height = 6)

# gpp <- rf_pdp1 %>%
#   ggplot(aes(x = mtry, y = Imp, color = Names,
#              group = Names, linetype = Names,
#              shape = Names)) +
#   geom_line() +
#   scale_linetype_manual(values = rep(2, each = 8)) +
#   scale_x_continuous(limits = c(1,8), breaks = seq(2,8,by=2)) +
#   scale_y_continuous(limits = c(0,max(rf_pdp1$Imp))) +
#   #scale_y_continuous(limits = c(0,4), breaks = seq(0,4,by=1)) +
#   ggtitle("PDP Variable Importance") +
#   geom_point(size = 5) +
#   scale_shape_manual(values = c(52,52,50,50,49,49,48,48)) +
#   scale_size(range = c(6,6)) +
#   ylab("Importance") +
#   guides(size = "none") +
#   theme(axis.text = element_text(size = 12),
#         axis.title = element_text(size = 15),
#         plot.title = element_text(size = 14, face = "bold")) +
#   easy_center_title() + easy_plot_legend_size(size = 11)
# gpp
# ggsave("xor_pdp_zoom.pdf", plot = last_plot(), dpi = 2400,
#        width = 6, height = 6)
```

```
library(patchwork)
gr + gp + gd + gv
```



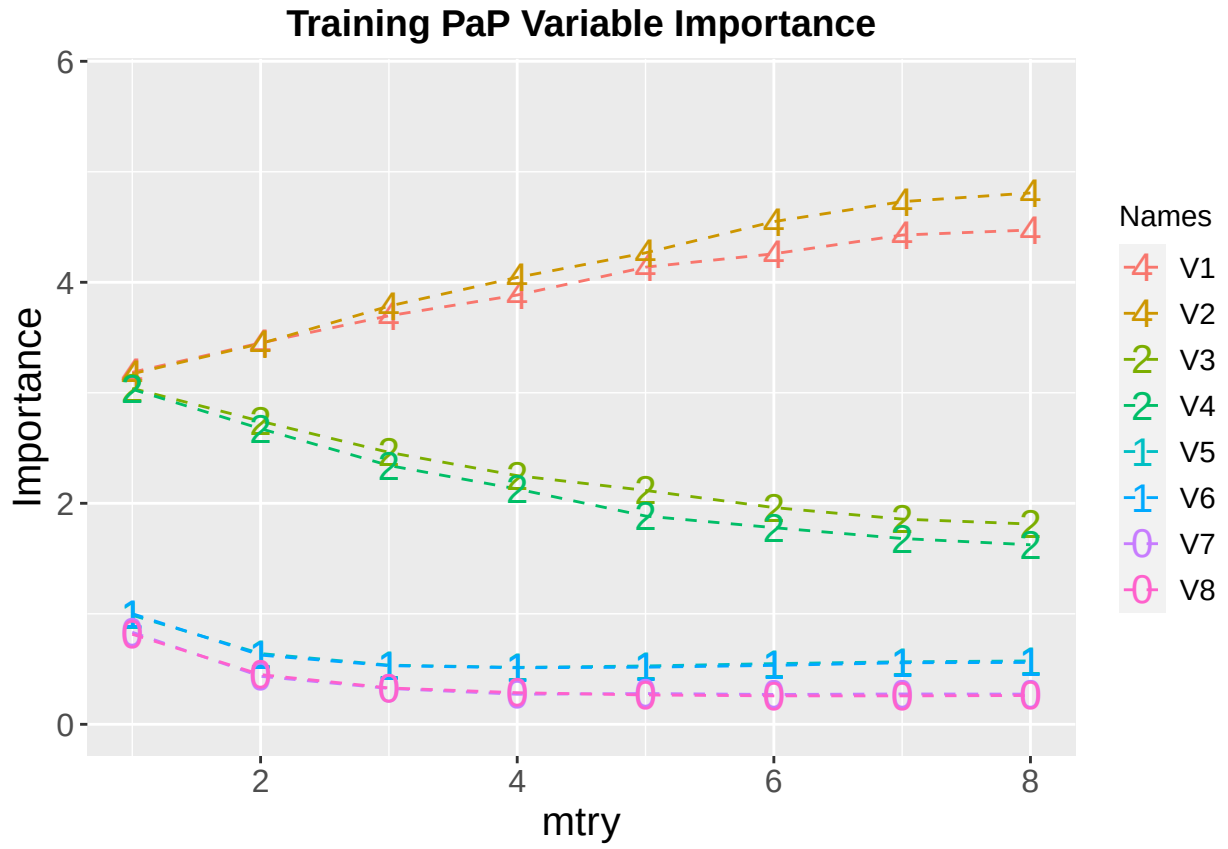

```
ggsave("xor_all_zoom.pdf", plot = gr + gp + gd + gv, dpi = 2400,
       width = 9, height = 9)
```

```
gr <- rf_oob_f1 %>%
  ggplot(aes(x = mtry, y = Imp, color = Names,
             group = Names, linetype = Names,
             shape = Names)) +
  geom_line() +
  scale_x_continuous(limits = c(1,8), breaks = seq(2,8,by=2)) +
  scale_y_continuous(limits = c(0,ma)) +
  ggtitle("OOB PaP Variable Importance") +
  geom_point(size = 5) +
  scale_linetype_manual(values = rep(2, each = 8)) +
  scale_shape_manual(values = c(52,52,50,50,49,49,48,48)) +
  scale_size(range = c(6,6)) +
  ylab("Importance") +
  guides(size = "none") +
  theme(axis.text = element_text(size = 12),
        axis.title = element_text(size = 15),
        plot.title = element_text(size = 14, face = "bold")) +
  easy_center_title() + easy_plot_legend_size(size = 11)
gr
```



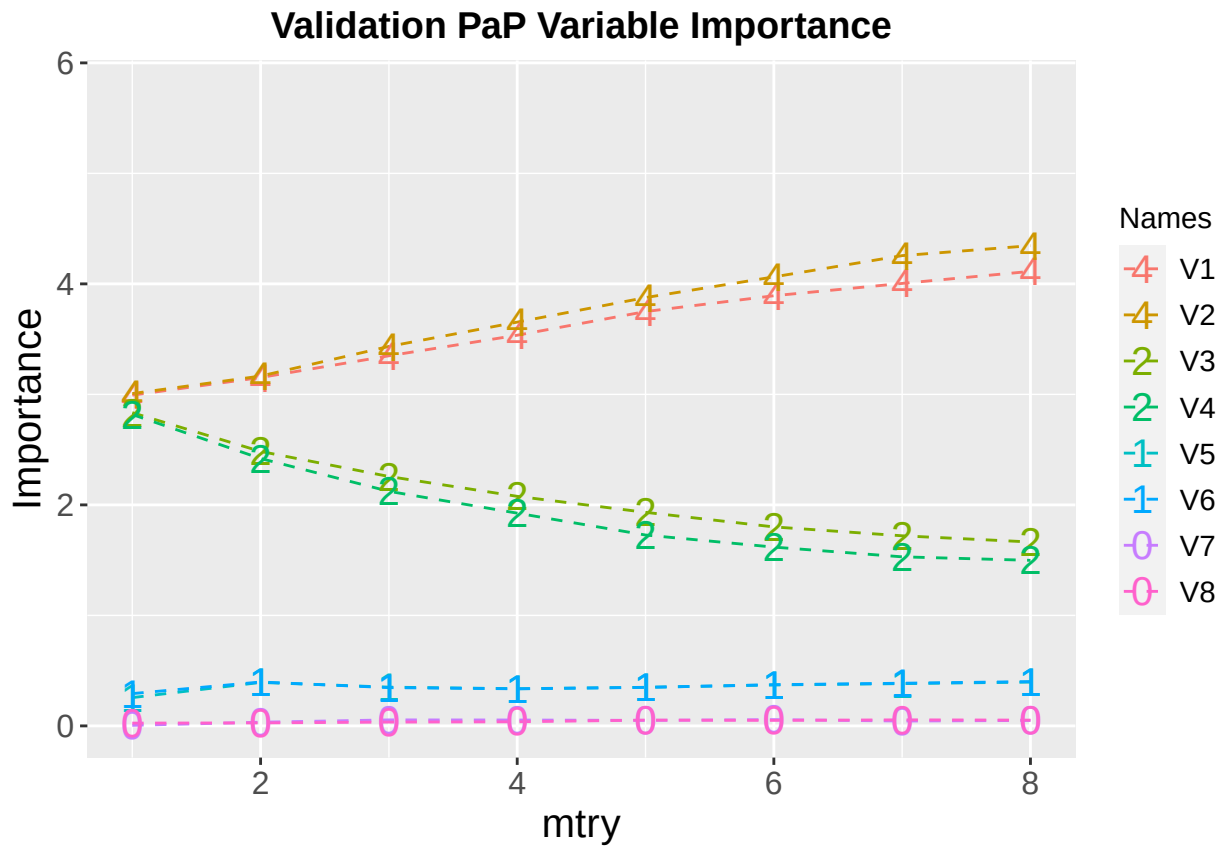

```
# ggsave("xor_oob.pdf", plot = last_plot(), dpi = 2400,
#        width = 6, height = 6)

gp <- perm_train1 %>%
  ggplot(aes(x = mtry, y = Imp, color = Names,
```

```
            group = Names, linetype = Names,
            shape = Names)) +
  geom_line() +
  scale_x_continuous(limits = c(1,8), breaks = seq(2,8,by=2)) +
  scale_y_continuous(limits = c(0,ma)) +
  ggtitle("Training PaP Variable Importance") +
  geom_point(size = 5) +
  scale_linetype_manual(values = rep(2, each = 8)) +
  scale_shape_manual(values = c(52,52,50,50,49,49,48,48)) +
  scale_size(range = c(6,6)) +
  ylab("Importance") +
  guides(size = "none") +
  theme(axis.text = element_text(size = 12),
        axis.title = element_text(size = 15),
        plot.title = element_text(size = 14, face = "bold")) +
  easy_center_title() + easy_plot_legend_size(size = 11)
gp
```



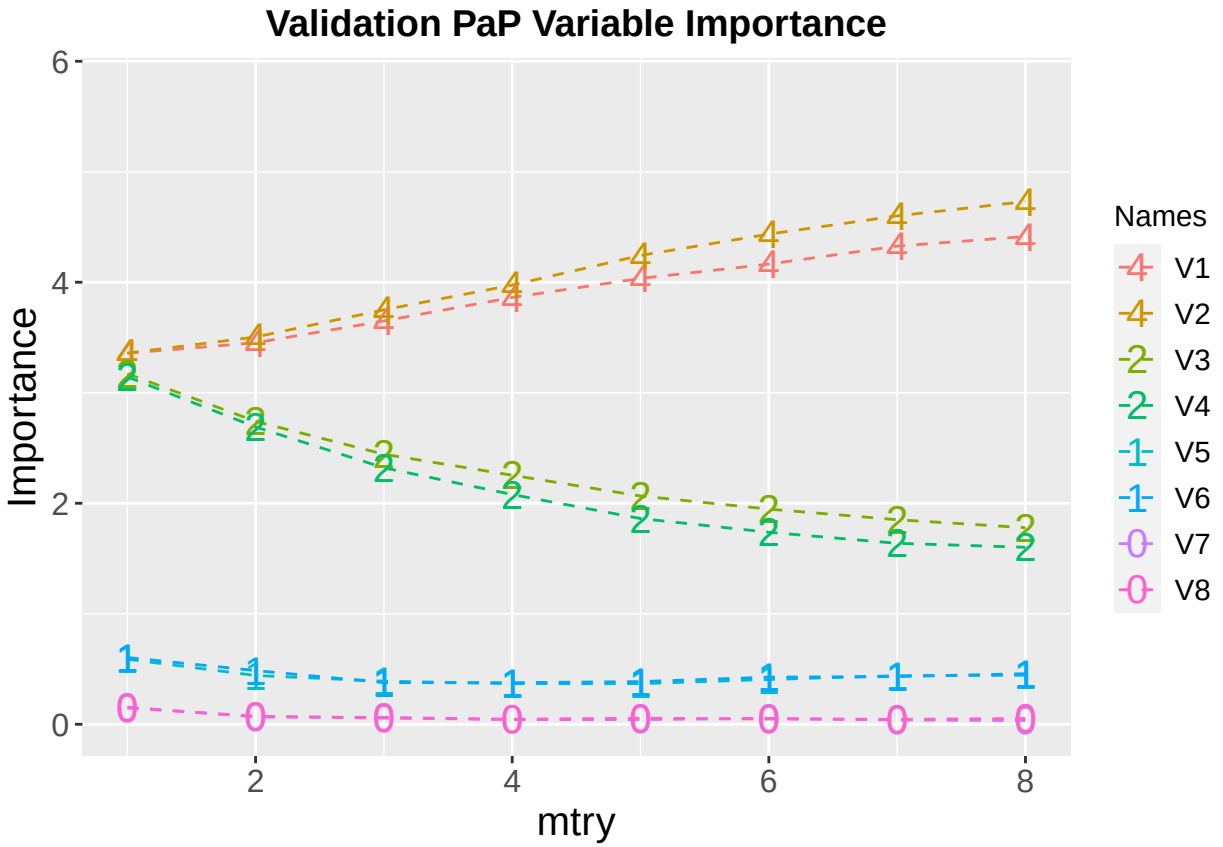

```
# ggsave("xor_train.pdf", plot = last_plot(), dpi = 2400,
#        width = 6, height = 6)

gd <- drop_valid1 %>%
  ggplot(aes(x = mtry, y = Imp, color = Names,
            group = Names, linetype = Names,
            shape = Names)) +
```

```
  geom_line() +
  scale_x_continuous(limits = c(1,8), breaks = seq(2,8,by=2)) +
  scale_y_continuous(limits = c(0,ma)) +
  ggtitle("Validation PaP Variable Importance") +
  geom_point(size = 5) +
  scale_linetype_manual(values = rep(2, each = 8)) +
  scale_shape_manual(values = c(52,52,50,50,49,49,48,48)) +
  scale_size(range = c(6,6)) +
  ylab("Importance") +
  guides(size = "none") +
  theme(axis.text = element_text(size = 12),
        axis.title = element_text(size = 15),
        plot.title = element_text(size = 14, face = "bold")) +
  easy_center_title() + easy_plot_legend_size(size = 11)
gd
```



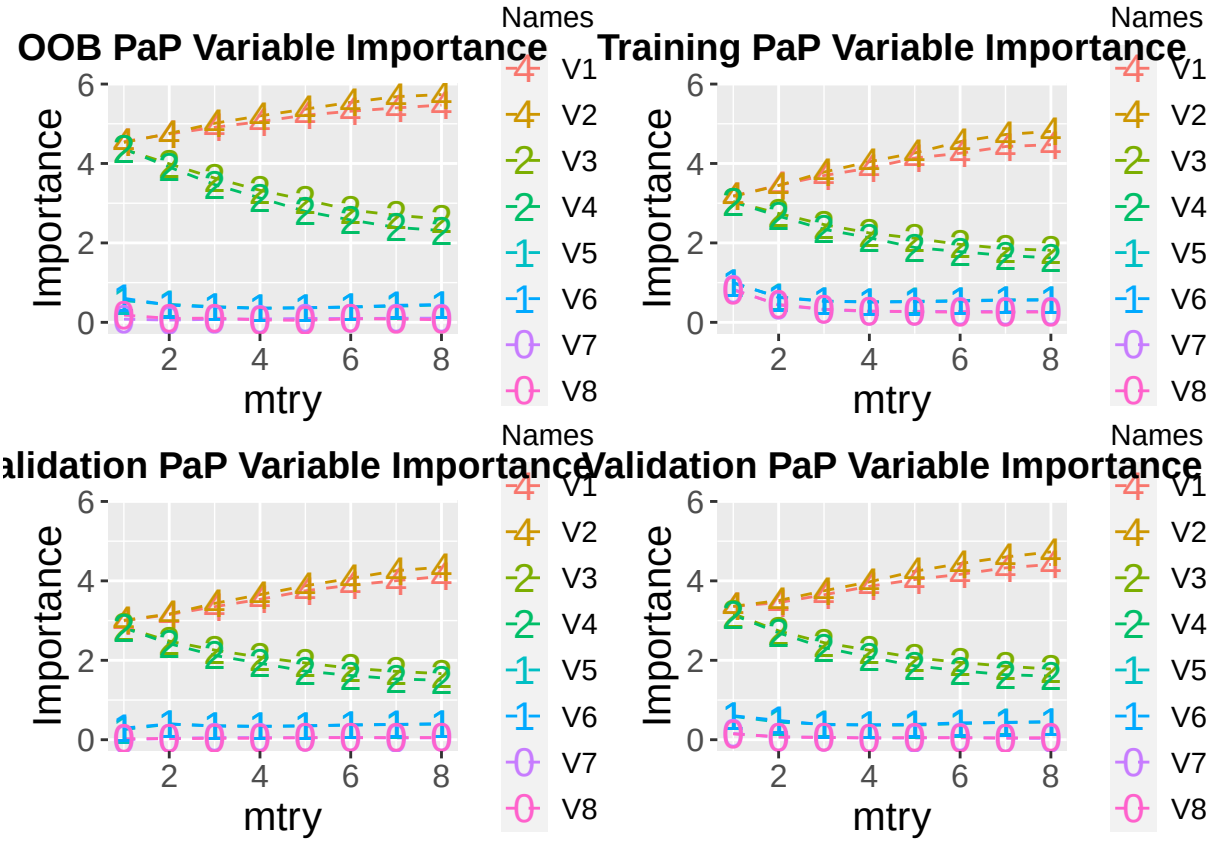

```
gv <- perm_valid1 %>%
  ggplot(aes(x = mtry, y = Imp, color = Names,
             group = Names, linetype = Names,
             shape = Names)) +
  geom_line() +
  scale_x_continuous(limits = c(1,8), breaks = seq(2,8,by=2)) +
  scale_y_continuous(limits = c(0,ma)) +
  ggtitle("Validation PaP Variable Importance") +
  geom_point(size = 5) +
```

```
  scale_linetype_manual(values = rep(2, each = 8)) +
  scale_shape_manual(values = c(52,52,50,50,49,49,48,48)) +
  scale_size(range = c(6,6)) +
  ylab("Importance") +
  guides(size = "none") +
  theme(axis.text = element_text(size = 12),
        axis.title = element_text(size = 15),
        plot.title = element_text(size = 14, face = "bold")) +
  easy_center_title() + easy_plot_legend_size(size = 11)
gv
```




```
# ggsave("xor_val.pdf", plot = last_plot(), dpi = 2400,
#        width = 6, height = 6)

# gpp <- rf_pdp1 %>%
#   ggplot(aes(x = mtry, y = Imp, color = Names,
#              group = Names, linetype = Names,
#              shape = Names)) +
#   geom_line() +
#   scale_linetype_manual(values = rep(2, each = 8)) +
#   scale_x_continuous(limits = c(1,8), breaks = seq(2,8,by=2)) +
#   scale_y_continuous(limits = c(0,mp)) +
#   ggtitle("PDP Variable Importance") +
#   geom_point(size = 5) +
#   scale_shape_manual(values = c(52,52,50,50,49,49,48,48)) +
```

```
#   scale_size(range = c(6,6)) +
#   ylab("Importance") +
#   guides(size = "none") +
#   theme(axis.text = element_text(size = 12),
#         axis.title = element_text(size = 15),
#         plot.title = element_text(size = 14, face = "bold")) +
#   easy_center_title() + easy_plot_legend_size(size = 11)
# gpp
# ggsave("xor_pdp.pdf", plot = last_plot(), dpi = 2400,
#        width = 6, height = 6)

library(patchwork)
gr + gp + gd + gv
```




```
ggsave("xor_all.pdf", plot = gr + gp + gd + gv, dpi = 2400,
       width = 9, height = 9)

Sys.time() - s
```

```
## Time difference of 11.47977 mins
```