XGBoost Comparison

Andrew vanderWilden

6/10/2020

```
library(tidyverse)
library(xgboost)
library(Metrics)

set.seed(5353)

df = read.table("auto-ins.tsv", header = TRUE)
```

```
df <- df %>%
  mutate(reg.grp = case_when(
      region == "Highlands" ~ "HLPC",
      region == "Lakeview" ~ "HLPC",
      region == "Piety Corner" ~ "HLPC",
      region == "Bleachery" ~ "BSBW",
      region == "Banks Square" ~ "BSBW",
      region == "Warrendale" ~ "BSBW",
      region == "The Lanes" ~ "HLPC",
      region == "The Chemistry" ~ "Chem",
      TRUE ~ "Other"))

df$reg.grp <- factor(df$reg.grp,
                     levels = c("HLPC", "BSBW", "Chem"))

df <- df %>%
    mutate(use.colps = case_when(
      vehicle.use == "Business" ~ "B&C",
      vehicle.use == "Commute" ~ "B&C",
      vehicle.use == "Private" ~ "Prv",
      TRUE ~ "Other"))

df$use.colps <- factor(df$use.colps,
                       levels = c("Prv", "B&C"))

df$v.bod <- fct_collapse(df$vehicle.body,
                         FML = c("Minibus", "Station Wagon"),
                         MED = c("Hatchback", "SUV", "Sedan"),
                         HST = c("Truck", "Panel Van", "Roadster"))

bks <- c(16, 20, 24, 30, 40, 50, 60, 76)
```

```
lbs <- c("17-20", "21-24", "25-30", "31-40",
         "41-50", "51-60", "61-75")

df$age.cat <- cut(df$age, breaks = bks, labels = lbs)
```

```
df1 <- df %>%
  select(claims,
         gender,
         age.cat,
         use.colps,
         reg.grp,
         v.bod,
         exposure)

head(df1)
```

```
##   claims gender age.cat use.colps reg.grp v.bod exposure
## 1      1 Female   17-20       Prv    HLPC   MED     1.00
## 2      0   Male   61-75       Prv    HLPC   MED     1.00
## 3      0 Female   31-40       B&C    HLPC   MED     0.75
## 4      0 Female   25-30       B&C    HLPC   MED     0.50
## 5      2   Male   21-24       B&C    Chem   MED     1.00
## 6      0   Male   17-20       B&C    HLPC   MED     1.00
```

```
Gender <- model.matrix(~gender-1, df1)
Age <- model.matrix(~age.cat-1, df1)
Use <- model.matrix(~use.colps-1, df1)
Reg <- model.matrix(~reg.grp-1, df1)
Body <- model.matrix(~v.bod-1, df1)


df_num <- cbind(Gender, Age, Use, Reg, Body)

df_matrix <- data.matrix(df_num)

df_label <- as_vector(df1$claims)
df_exposure <- as_vector(df1$exposure)

train_idx <- sample(10000,7000, replace = FALSE)

train_data <- df_matrix[train_idx,]
train_label <- df_label[train_idx]
train_exposure <- df_exposure[train_idx]

test_data <- df_matrix[-train_idx,]
test_label <- df_label[-train_idx]
test_exposure <- df_exposure[-train_idx]

dtrain <- xgb.DMatrix(data = train_data, label = train_label)
dtest <- xgb.DMatrix(data = test_data, label = test_label)

setinfo(dtrain, "base_margin", log(train_exposure)) # For offset
```

```
## [1] TRUE
```

```
setinfo(dtest, "base_margin", log(test_exposure))
```

```
## [1] TRUE
```

```
fin_mod <- glm(claims~gender + age.cat + use.colps + reg.grp + v.bod + gender:age.cat, data = df1[train_
```

```
asdf <- fin_mod$coefficients

fdsa <- tibble(Coefficient = names(asdf), Value = asdf)

knitr::kable(asdf)
```

	x
(Intercept)	-2.2058848
genderMale	0.4904278
age.cat21-24	-0.1678105
age.cat25-30	-0.3144790
age.cat31-40	-0.6812966
age.cat41-50	-0.8779139
age.cat51-60	-1.7584890
age.cat61-75	-0.3582758
use.colpsB&C	0.7022586
reg.grpBSBW	-0.3831956
reg.grpChem	-0.2166617
v.bodFML	-0.4413887
v.bodHST	0.2789719
genderMale:age.cat21-24	-0.0887434
genderMale:age.cat25-30	-0.0514458
genderMale:age.cat31-40	-0.1558785
genderMale:age.cat41-50	-0.0157540
genderMale:age.cat51-60	0.6719896
genderMale:age.cat61-75	-0.8294848

```
model <- xgboost(data = dtrain,
                 objective = "count:poisson",
                 max.depth = 3,
                 early_stopping_rounds = 4,
                 print_every_n = 10,
                 nrounds = 500,
                 min_child_weight = 1,
                 gamma = 1)
```

```
## [1]  train-poisson-nloglik:0.822982
## Will train until train_poisson_nloglik hasn't improved in 4 rounds.
##
## [11] train-poisson-nloglik:0.401141
## [21] train-poisson-nloglik:0.340956
```

```
## [31] train-poisson-nloglik:0.332176
## [41] train-poisson-nloglik:0.329866
## [51] train-poisson-nloglik:0.329513
## Stopping. Best iteration:
## [51] train-poisson-nloglik:0.329513
```

```
model2 <- xgboost(data = dtrain,
                  objective = "count:poisson",
                  max.depth = 2,
                  early_stopping_rounds = 4,
                  print_every_n = 10,
                  nrounds = 500,
                  min_child_weight = 1,
                  gamma = 1)
```

```
## [1]  train-poisson-nloglik:0.823034
## Will train until train_poisson_nloglik hasn't improved in 4 rounds.
##
## [11] train-poisson-nloglik:0.402943
## [21] train-poisson-nloglik:0.344469
## [31] train-poisson-nloglik:0.336040
## [41] train-poisson-nloglik:0.333399
## [51] train-poisson-nloglik:0.331704
## Stopping. Best iteration:
## [56] train-poisson-nloglik:0.331222
```

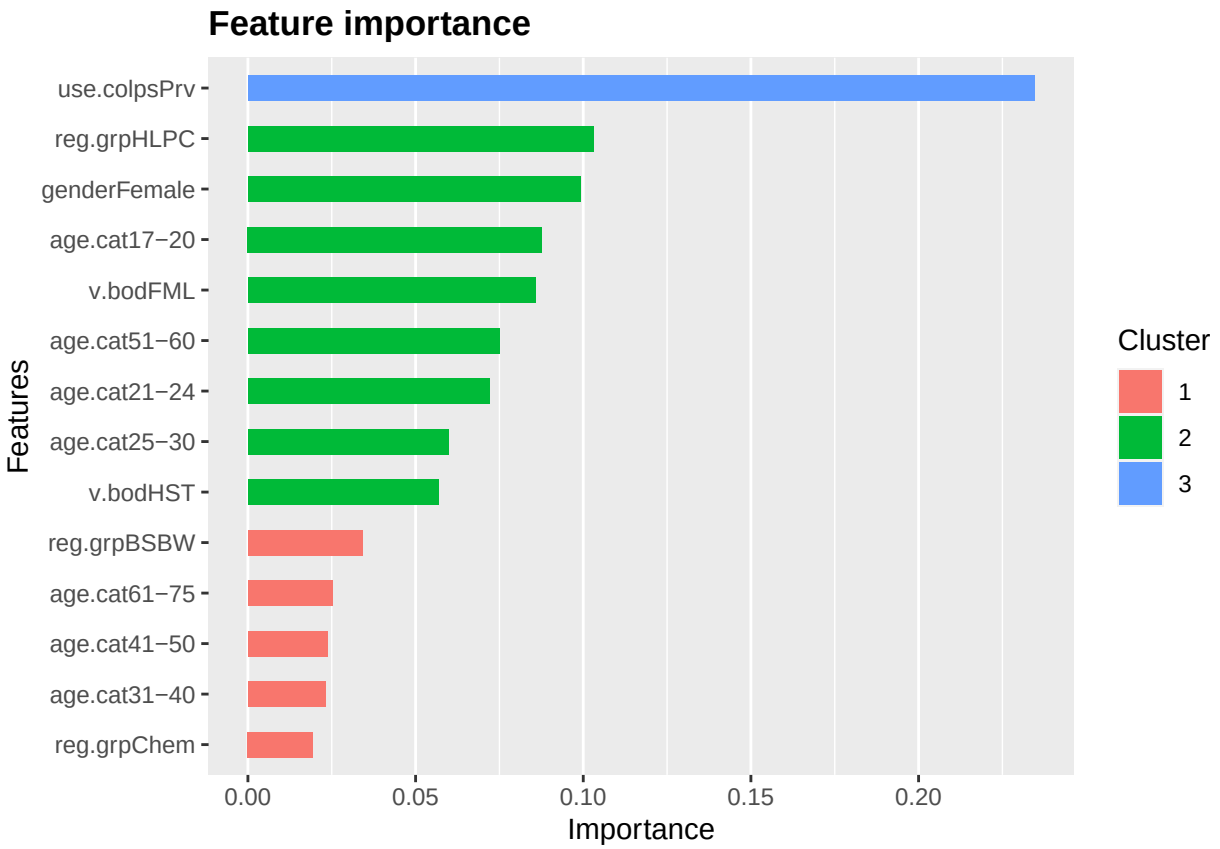
```
model3 <- xgboost(data = dtrain,
                  objective = "count:poisson",
                  max.depth = 2,
                  #early_stopping_rounds = 4,
                  print_every_n = 10,
                  nrounds = 500,
                  min_child_weight = 1,
                  gamma = 1)
```

```
## [1]  train-poisson-nloglik:0.823034
## [11] train-poisson-nloglik:0.402943
## [21] train-poisson-nloglik:0.344469
## [31] train-poisson-nloglik:0.336040
## [41] train-poisson-nloglik:0.333399
## [51] train-poisson-nloglik:0.331704
## [61] train-poisson-nloglik:0.331222
## [71] train-poisson-nloglik:0.331222
## [81] train-poisson-nloglik:0.331222
## [91] train-poisson-nloglik:0.331222
## [101]    train-poisson-nloglik:0.331222
## [111]    train-poisson-nloglik:0.331222
## [121]    train-poisson-nloglik:0.331222
## [131]    train-poisson-nloglik:0.331222
## [141]    train-poisson-nloglik:0.331222
## [151]    train-poisson-nloglik:0.331222
## [161]    train-poisson-nloglik:0.331222
## [171]    train-poisson-nloglik:0.331222
```

```
## [181]    train-poisson-nloglik:0.331222
## [191]    train-poisson-nloglik:0.331222
## [201]    train-poisson-nloglik:0.331222
## [211]    train-poisson-nloglik:0.331222
## [221]    train-poisson-nloglik:0.331222
## [231]    train-poisson-nloglik:0.331222
## [241]    train-poisson-nloglik:0.331222
## [251]    train-poisson-nloglik:0.331222
## [261]    train-poisson-nloglik:0.331222
## [271]    train-poisson-nloglik:0.331222
## [281]    train-poisson-nloglik:0.331222
## [291]    train-poisson-nloglik:0.331222
## [301]    train-poisson-nloglik:0.331222
## [311]    train-poisson-nloglik:0.331222
## [321]    train-poisson-nloglik:0.331222
## [331]    train-poisson-nloglik:0.331222
## [341]    train-poisson-nloglik:0.331222
## [351]    train-poisson-nloglik:0.331222
## [361]    train-poisson-nloglik:0.331222
## [371]    train-poisson-nloglik:0.331222
## [381]    train-poisson-nloglik:0.331222
## [391]    train-poisson-nloglik:0.331222
## [401]    train-poisson-nloglik:0.331222
## [411]    train-poisson-nloglik:0.331222
## [421]    train-poisson-nloglik:0.331222
## [431]    train-poisson-nloglik:0.331222
## [441]    train-poisson-nloglik:0.331222
## [451]    train-poisson-nloglik:0.331222
## [461]    train-poisson-nloglik:0.331222
## [471]    train-poisson-nloglik:0.331222
## [481]    train-poisson-nloglik:0.331222
## [491]    train-poisson-nloglik:0.331222
## [500]    train-poisson-nloglik:0.331222
```

```
mat <- xgb.importance(names(df_matrix),
                      model = model)

xgb.ggplot.importance(mat)
```




```
xgb_pred <- predict(model, dtest)
xgb2_pred <- predict(model2, dtest)
xgb3_pred <- predict(model3, dtest)
glm_pred <- predict(fin_mod, df1[-train_idx,], type = "response")

xgb2_o_pred <- predict(model2, dtrain)
glm_o_pred <- predict(fin_mod, df1[train_idx,], type = "response")
xgb3_o_pred <- predict(model3, dtrain)
```

```
msep_fit <- function(x,y) {
  ans <- mean((y-x)^2)
  return(ans)
}
```

```
msep_fit(xgb_pred, test_label)
```

```
## [1] 0.1172858
```

```
msep_fit(xgb2_pred, test_label)
```

```
## [1] 0.117169
```

```
msep_fit(xgb3_pred, test_label)
```

```
## [1] 0.1171656
```

```
msep_fit(glm_pred, test_label)
```

```
## [1] 0.1172515
```

```
msep_fit(xgb2_o_pred, train_label)
```

```
## [1] 0.1370088
```

```
msep_fit(xgb3_o_pred, train_label)
```

```
## [1] 0.1370118
```

```
msep_fit(glm_o_pred, train_label)
```

```
## [1] 0.1372614
```