DMC@ISU: The 2015 Iowa State University Data Mining Cup Team

Curation and Cross Validation to Reduce Features

Spring 2015, A Team as Strong as Steel

Last Day: May 19, 2015

I am using the following packages:

```
   library(magrittr)
   library(dplyr)
   library(reshape2)
   library(tidyr)
   library(lubridate)
   library(ggplot2)
   library(gridExtra)
   library(directlabels)
   library(rCharts)
   library(xtable)
   library(foreach)
   library(gtools)
   library(knitr)
   library(utils)
   library(GGally)
   source("~/dmc2015/ian/R/renm.R")

predPlot = function(pred_set1, pred_set2) {
    p = qplot(pred_set1, pred_set2) + coord_fixed() + geom_abline(intercept = 0,
        slope = 1, color = "red")
    print(p)
    return(p)
}

lossFunction = function(pred, actual) sum(((1/mean(actual)) * (pred - actual))^2)

predictor_weight = function(pred1, pred2, actual) {
    alphas = seq(0, 1, 0.001)
    err = sapply(alphas, function(alpha) lossFunction(alpha * pred1 + (1 - alpha) *
        pred2, actual))
    pred = data.frame(proportion.1 = alphas, err = err)
    return(pred)
}
```

My working directory is set to `~/dmc2015/predictions/`.

0.1 Set 1

I am using our new clean data - so should you

```
source("~/dmc2015/ian/load_data.r")

## using the following as id:
##  orderID,
##  orderTime,
```

```
##  userID,
##  couponsReceived,
##  basketValue,
##  couponsReceivedDate,
##  couponsReceivedTime,
##  couponsReceivedDoW,
##  couponsReceivedWeekend,
##  couponsReceivedFriSat,
##  orderTimeDate,
##  orderTimeTime,
##  orderTimeDoW,
##  orderTimeWeekend,
##  orderTimeFriSat,
##  batchID,
##  couponsExpire,
##  couponsSent,
##  TimeBtwnSentRec,
##  TimeBtwnRecExpire,
##  TimeBtwnRecOrder,
##  TimeBtwnOrderExpire,
##  ShopFast,
##  EarlyRec,
##  Shop60,
##  Shop30,
##  Shop15,
##  RecExpire60,
##  RecOrder60,
##  OrderExpire60
##
## using the following as measure columns:
##  couponID1,
##  price1,
##  basePrice1,
##  reward1,
##  premiumProduct1,
##  brand1,
##  productGroup1,
##  categoryIDs1,
##  coupon1Used,
##  basePrice_price_ratio1,
##  couponID2,
##  price2,
##  basePrice2,
##  reward2,
##  premiumProduct2,
##  brand2,
##  productGroup2,
##  categoryIDs2,
##  coupon2Used,
##  basePrice_price_ratio2,
##  couponID3,
##  price3,
##  basePrice3,
##  reward3,
```

```
##  premiumProduct3,
##  brand3,
##  productGroup3,
##  categoryIDs3,
##  coupon3Used,
##  basePrice_price_ratio3

HTVset1 = readRDS("~/dmc2015/data/featureMatrix/HTVset1.rds")
actual.val = as.vector(t(HTVset1$V[, c("coupon1Used", "coupon2Used", "coupon3Used")]))
```

0.1.1 Reading in the predictions

The predictions are stored in `~/dmc2015/predictions/`. We can read the files into R using the `list.files` function.

```
files = list.files("~/dmc2015/predictions/set1/", pattern = "rds", full.names = TRUE)
files
```

```
##  [1] "/Users/user/dmc2015/predictions/set1//crf_386col_set1_0.8.rds"
##  [2] "/Users/user/dmc2015/predictions/set1//crf_ada_set1_0.3.rds"
##  [3] "/Users/user/dmc2015/predictions/set1//crf_c50_set1_0.3.rds"
##  [4] "/Users/user/dmc2015/predictions/set1//crf_c50_set1_0.4.rds"
##  [5] "/Users/user/dmc2015/predictions/set1//crf_crf_set1_0.3.rds"
##  [6] "/Users/user/dmc2015/predictions/set1//crf_crf_set1_0.5.rds"
##  [7] "/Users/user/dmc2015/predictions/set1//crf_lasso_set1_0.3.rds"
##  [8] "/Users/user/dmc2015/predictions/set1//crf_lasso_set1_0.4.rds"
##  [9] "/Users/user/dmc2015/predictions/set1//crf_rf_set1_0.3.rds"
## [10] "/Users/user/dmc2015/predictions/set1//crf_rf_set1_0.4.rds"
## [11] "/Users/user/dmc2015/predictions/set1//gbm_386col_set1_0.8.rds"
## [12] "/Users/user/dmc2015/predictions/set1//gbm_rf_set1_0.8.rds"
## [13] "/Users/user/dmc2015/predictions/set1//lasso_386col_set1.rds"
## [14] "/Users/user/dmc2015/predictions/set1//lasso_c50_set1.rds"
## [15] "/Users/user/dmc2015/predictions/set1//lasso_crf_set1.rds"
## [16] "/Users/user/dmc2015/predictions/set1//lasso_gbm_set1.rds"
## [17] "/Users/user/dmc2015/predictions/set1//lasso_lasso_set1.rds"
## [18] "/Users/user/dmc2015/predictions/set1//lasso_rf_set1.rds"
## [19] "/Users/user/dmc2015/predictions/set1//pca_386col_set1.rds"
## [20] "/Users/user/dmc2015/predictions/set1//pca_c50_set1.rds"
## [21] "/Users/user/dmc2015/predictions/set1//pca_crf_set1.rds"
## [22] "/Users/user/dmc2015/predictions/set1//pca_gbm_set1.rds"
## [23] "/Users/user/dmc2015/predictions/set1//pca_lasso_set1.rds"
## [24] "/Users/user/dmc2015/predictions/set1//pca_rf_set1.rds"
## [25] "/Users/user/dmc2015/predictions/set1//rf_386col_set1.rds"
## [26] "/Users/user/dmc2015/predictions/set1//rf_c50_set1.rds"
## [27] "/Users/user/dmc2015/predictions/set1//rf_lasso_set1.rds"
## [28] "/Users/user/dmc2015/predictions/set1//rf_rf_set1.rds"
```

```
# base names
base_names = gsub("(/Users/user/dmc2015/predictions/set1//)(.*)(\\.rds)", "\\2",
    files)
```

Many of the sets have the **DMC Official** structure:

```
DMCofficial = readRDS("~/dmc2015/predictions/set1/crf_386col_set1_0.8.rds")
str(DMCofficial)
```

```
## List of 4
##  $ val_predictions  : num [1:4035, 1] 0.208 0.245 0.217 0.252 0.193 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr "couponUsed"
##  $ class_predictions: num [1:2007, 1:2] 6054 6054 6054 6055 6055 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:2] "orderID" "couponUsed"
##  $ error            : num [1:4] 3931 5985 6463 16379
##  $ details          :List of 4
##   ..$ vars  : chr "Peng Liahua's variables"
##   ..$ nvars : num 386
##   ..$ ntrees: num 1000
##   ..$ mtry  : num 10
```

I can check the structure in a similar way to how I read in features:

```
preds = lapply(files, function(x) readRDS(x))
pred.struc = rep("renegade style", length(files))

for (i in 1:length(files)) {
    if (all(names(preds[[i]]) %in% names(DMCofficial))) {
        pred.struc[i] = "DMC Official"
    }
}
```

```
pred.struc
```

```
##  [1] "DMC Official"   "DMC Official"   "DMC Official"   "DMC Official"
##  [5] "DMC Official"   "DMC Official"   "DMC Official"   "DMC Official"
##  [9] "DMC Official"   "DMC Official"   "renegade style" "renegade style"
## [13] "renegade style" "renegade style" "renegade style" "renegade style"
## [17] "renegade style" "renegade style" "renegade style" "renegade style"
## [21] "renegade style" "renegade style" "renegade style" "renegade style"
## [25] "renegade style" "renegade style" "renegade style" "renegade style"
```

We can get the DMC official **validation predictions** like this:

```
# which files were DMC Official?
these_ones = which(pred.struc == "DMC Official")

# read the set
validation_predictions = HTVset1$V %>% select(orderID, coupon1Used, coupon2Used,
    coupon3Used) %>% gather(colname, couponUsed, -orderID) %>% mutate(couponCol = as.numeric(gsub("coup
    "\\1", colname))) %>% select(orderID, couponCol, couponUsed) %>% arrange(orderID,
    couponCol)

# make do.call dplyr friendly
you.call = function(x, func) do.call(func, x)

# extract validation predictions
pred.val = these_ones %>% lapply(function(i) preds[[i]]$val_predictions) %>%
    you.call("cbind") %>% data.frame
```

```
names(pred.val) = base_names[these_ones]

validation_predictions = validation_predictions %>% cbind(pred.val)
```

and the DMC Official classification predictions like this:

```
classification_predictions = HTVset1$C %>% select(orderID, coupon1Used, coupon2Used,
    coupon3Used) %>% gather(colname, couponUsed, -orderID) %>% mutate(couponCol = as.numeric(gsub("coup
    "\\1", colname))) %>% select(orderID, couponCol, couponUsed) %>% arrange(orderID,
    couponCol)

# extract validation predictions
pred.class = these_ones %>% lapply(function(i) preds[[i]]$class_predictions[,
    2]) %>% you.call("cbind") %>% data.frame

# base names
base_names = gsub("(/Users/user/dmc2015/predictions/set1//)(.*)(\\.rds)", "\\2",
    files)

names(pred.class) = base_names[these_ones]

classification_predictions = classification_predictions %>% cbind(pred.class)
```

The rest of the files follow the same arrangement:

```
# get the validation predictions:
for (i in 11:18) {
    validation_predictions = validation_predictions %>% left_join(preds[[i]]$validation,
        by = c("orderID", "couponCol"))

    # get the validation predictions:
    classification_predictions = classification_predictions %>% left_join(preds[[i]]$class,
        by = c("orderID", "couponCol"))
}
```

0.1.2 Prediction Geometry

Let's get the plots:

```
ggpairs(validation_predictions, 4:21)
```