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

Initial Exploration

Spring 2015, Iowa State

Due Date: April 14 2015

I am using the following packages:

```
library(ggplot2)
library(lubridate)
library(xtable)
library(foreach)
library(rCharts)
library(plyr)
library(dplyr)
library(reshape2)
library(gtools)
library(sqldf)
```

and my working directory is set to `dmc2015/ian`.

0.1 Reading the Data

Read the data into R:

```
# training set ('historical data')
trn.raw <- read.delim("../data/raw_data/DMC_2015_orders_train.txt",
    stringsAsFactors = FALSE, sep = "|", quote = "")
trn <- read.csv("../data/clean_data/train_simple_name.csv",
    stringsAsFactors = FALSE, na.strings = "")

# test set ('future data')
tst.raw <- read.delim("../data/raw_data/DMC_2015_orders_class.txt",
    stringsAsFactors = FALSE, sep = "|", quote = "")
tst <- read.csv("../data/clean_data/test_simple_name.csv",
    stringsAsFactors = FALSE, na.strings = "")
```

add the time features:

```
source("~/dmc2015/ian/R/TimeFeatures.R")

# Whatever you do to the training set
trn <- TimeFeatures(trn, "orderTime")
trn <- TimeFeatures(trn, "couponsReceived")

# try if you can to do the same to the test set
tst <- TimeFeatures(tst, "orderTime")
tst <- TimeFeatures(tst, "couponsReceived")
```

and identify orders as belonging to batches starting at every Tuesday at midnight and lasting for one week.

```
# add batch information
source("~/dmc2015/ian/R/GetBatchInfo.R")
batchres <- GetBatchInfo("2015-01-06 00:00:01", unts = "hours")

trn <- batchres$train
tst <- batchres$test

# write the batchID and couponsSent as features:
trainFeatures <- trn[, c("orderID", names(trn)[!(names(trn) %in%
    names(trn.raw))])]

write.csv(trainFeatures, file = "~/dmc2015/features/feature_files/batchInfo_train.csv",
    row.names = FALSE, na = "", quote = FALSE)

testFeatures <- tst[, c("orderID", names(tst)[!(names(tst) %in%
    names(tst.raw))])]

write.csv(testFeatures, file = "~/dmc2015/features/feature_files/batchInfo_test.csv",
    row.names = FALSE, na = "", quote = FALSE)
```

Because some of these values are formatted, I also have created the following `rds` files which can be read using `readRDS`.

```
saveRDS(trainFeatures, "~/dmc2015/features/feature_files/batchInfo_train.rds")
saveRDS(testFeatures, "~/dmc2015/features/feature_files/batchInfo_test.rds")
```