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(directlabels)
library(rCharts)
library(xtable)
library(foreach)
library(gtools)
library(knitr)
library(utils)
source("~/dmc2015/ian/R/renm.R")
```

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

0.1 Pete Check

Pete's in MN right now by he's soldiering through:

0.1.1 Reading in Pete's Work

A few notes - there is definetly an interesting set of variables in this mix (particularly set 2's variables). I

```
PeteImp = readRDS("../pete/predictions/importance.rds")

## 20 most important features in Set 1
PeteImp %>% arrange(-h1_imp) %>% select(var, h1_imp) %>% head(10)
```

```
##                                    var   h1_imp
## 1                  avgTime_recd_order 165.4315
## 2                TimeBtwnRecOrder_disc 162.2748
## 3  couponsReceivedDoW_orderTimeDoW_prob 156.8312
## 4               TimeBtwnRecOrder.median 153.4902
## 5            TimeBtwnRecOrder_disc_prob 150.6548
## 6      TimeBtwnRecOrder_disc_brand_prob 136.4971
## 7               llr_est_catidsXShopFast 131.0473
## 8       TimeBtwnRecOrder_disc_nUserUsed 130.5723
## 9                  llr_naive_RecOrder60 127.3065
## 10 llr_est_luxXcatidsXShopFastXEarlyRec 125.6800
```

```
## 20 most important features in Set 2
PeteImp %>% arrange(-h2_imp) %>% select(var, h2_imp) %>% head(10)
```

```
##                                      var   h2_imp
## 1                 TimeBtwnRecOrder.mean 154.9323
## 2                    avgTime_recd_order 151.5438
## 3                  TimeBtwnRecOrder.max 145.7172
## 4       TimeBtwnRecOrder_disc_nUserUsed 133.7729
## 5                 TimeBtwnRecOrder_disc 133.6563
## 6                    llr_est_RecOrder60 128.9762
## 7               TimeBtwnRecOrder.median 118.1619
## 8            TimeBtwnRecOrder_disc_prob 117.1151
## 9           llr_est_luxXcatidsXShopFast 116.7540
## 10 couponsReceivedDoW_orderTimeDoW_prob 115.4323
```

```
## 20 most important features in Set 3
PeteImp %>% arrange(-h3_imp) %>% select(var, h3_imp) %>% head(10)
```

```
##                                              var   h3_imp
## 1                            avgTime_recd_order 210.7144
## 2           couponsReceivedDoW_orderTimeDoW_prob 194.5691
## 3                     TimeBtwnRecOrder_disc_prob 183.4576
## 4                TimeBtwnRecOrder_disc_nUserUsed 178.8233
## 5                            llr_est_RecOrder60 172.1656
## 6        orderTimeDoW_TimeBtwnRecOrder_disc_prob 171.9130
## 7                          TimeBtwnRecOrder.max 161.6381
## 8  couponsReceivedDoW_TimeBtwnRecOrder_disc_prob 155.1606
## 9                         TimeBtwnRecOrder_disc 143.1713
## 10             reward_TimeBtwnRecOrder_disc_prob 142.3068
```

Alex is adapting these importance levels have been adapted to our variable selection set. Thanks Pete!