model_HA_DLag.R

atchirc

Tue May 23 00:46:14 2017

```
library(MASS)
library(car)
library(DataCombine)   # Pair wise correlation
library(stargazer)
library(dplyr)         # Data aggregation
library(glmnet)
source('../atchircUtils.R')


data    <- read.csv('../../intrim/eleckart.csv')


# KPI selection
# units, product_mrp, list_mrp, COD, Prepaid are factors
# Insig : Affiliates corr OnlineMarketing
# Insig : Radio  corr Other
# Insig : Digitial, ContentMarketing corr SEM
# delivery(b/c)days are corr, lets choose deliverycdays
# will use marketing levers rather TotalInvestment

# Filter significant KPIs
model_data <- subset(data, product_analytic_sub_category=='HomeAudio',
                     select = -c(product_analytic_sub_category,product_mrp,
                                 units,COD,Prepaid,deliverybdays,
                                 TotalInvestment,Affiliates,Radio,Digital,
                                 ContentMarketing,sla,procurement_sla))

model_data_org <- model_data
model_data[,c(8:12)] <- model_data[,c(8:12)]*10000000


# # ***************************************************************************
# #                   FEATURE ENGINEERING -PASS2  ----
# # ***************************************************************************
#
# # . . . . List Price Inflation ----
model_data$chnglist <- c(0,diff(model_data$list_mrp))
#
# # . . . . Discount Inflation ----
model_data$chngdisc <- c(0,diff(model_data$discount))
#

# # . . . . Lag independant variables----
# # Lag weekly avg discount by 1 week
model_data$laggmv       <- data.table::shift(model_data$gmv)
model_data$lagdiscount  <- data.table::shift(model_data$discount)
model_data$lagdeliverycdays <- data.table::shift(model_data$deliverycdays)
```

```
model_data$lagTV        <- data.table::shift(model_data$TV)
model_data$lagSponsorship <- data.table::shift(model_data$Sponsorship)
model_data$lagOnlineMar   <- data.table::shift(model_data$OnlineMarketing)
model_data$lagSEM         <- data.table::shift(model_data$SEM)
model_data$lagOther       <- data.table::shift(model_data$Other)
model_data$lagNPS         <- data.table::shift(model_data$NPS)
model_data$laglist_mrp    <- data.table::shift(model_data$list_mrp)
model_data$lagChnglist    <- data.table::shift(model_data$chnglist)
model_data$lagChngdisc    <- data.table::shift(model_data$chngdisc)
```

*

```
**PROCs:**
```

Linear, Ridge and Lasso Model are wrapped with abstract functions. This would facilitate readable code for model building and Model otpimization. Set Class definitions

```r
setOldClass('elnet')
setClass(Class = 'atcglmnet',
         representation (
           R2 = 'numeric',
           mdl = 'elnet',
           pred = 'matrix'
         )
)
```

```r
setOldClass('lm')
setClass(Class = 'atclm',
         representation (
           R2 = 'numeric',
           mdl = 'lm',
           pred = 'matrix'
         )
)
```

Finding min lambda from 1000 iterations Function to find Min Lambda using bootstrap method. minlambda identified over 1000 cross validation trails. observed minlambda used for Ridge and Lasso regression.

```r
findMinLambda <- function(x,y,alpha,folds) {
  lambda_list <- list()
  for (i in 1:1000) {
    cv.out <- cv.glmnet(as.matrix(x), as.vector(y), alpha=alpha,
                        nfolds=folds)
    lambda_list <- append(lambda_list, cv.out$lambda.min)
  }
  return(min(unlist(lambda_list)))
}
```

Linear Model with Regularization Wrapper function for Ridge and Lasso regression. functions performs Ridge/Lasso regression and returns R2, Model and Predicted values as `atcglmnet` object

```r
atcLmReg <- function(x,y,l1l2,folds) {
  # l1l2 = 0 for L1,  1 for L2

  if (l1l2) { # Lasso/L2
    min_lambda <- findMinLambda(x,y,1,folds)
  } else { # Ridge/L1
    min_lambda <- findMinLambda(x,y,0,folds)
  }
  mdl        <- glmnet(x,y,alpha=l1l2,lambda = min_lambda)
  pred       <- predict(mdl,s= min_lambda,newx=x)

  # MSE
  mean((pred-y)^2)
  R2 <- 1 - (sum((y-pred )^2)/sum((y-mean(pred))^2))
  return(new('atcglmnet', R2 = R2, mdl=mdl, pred=pred))
}
```

*

MODELING

```r
# Prune KPI as part of model optimization
model_data <- na.omit(model_data)
model_data <- subset(model_data, select=-c(list_mrp,laglist_mrp,
                                           TV,lagTV,NPS,lagNPS,discount,
                                           lagdiscount,OnlineMarketing,
                                           laggmv,deliverycdays,lagdeliverycdays,
                                           SEM,lagChnglist))
```

Linear Model:

```r
mdl      <- lm(gmv~., data=model_data)
step_mdl <- stepAIC(mdl,direction = 'both',trace = FALSE)

stargazer(mdl,step_mdl, align = TRUE, type = 'text',
          title='Linear Regression Results', single.row=TRUE)
```

```
##
## Linear Regression Results
## =================================================================================
##                                          Dependent variable:
##                     -------------------------------------------------------------
##                                                  gmv
##                                  (1)                            (2)
## ---------------------------------------------------------------------------------
## week                 -63,853.750*** (22,877.990)    -71,623.480*** (19,730.630)
## n_saledays            338,502.900* (175,654.900)     327,624.500* (167,371.500)
## Sponsorship            99,051.790 (87,301.360)       100,378.800* (51,763.810)
## Other                       0.008 (0.023)
## chnglist                   0.0003* (0.0002)              0.0003** (0.0002)
## chngdisc             211,894.800*** (50,523.780)    214,910.100*** (48,037.850)
## lagSponsorship         2,509.796 (114,363.700)
## lagOnlineMar                0.029 (0.024)                 0.043** (0.018)
## lagSEM                      0.017 (0.028)
## lagOther                    0.005 (0.024)
## lagChngdisc            81,462.560 (48,342.900)        78,069.660* (46,071.400)
## Constant            4,710,177.000*** (848,995.100) 4,806,894.000*** (801,606.400)
## ---------------------------------------------------------------------------------
## Observations                      49                             49
## R2                              0.653                          0.643
## Adjusted R2                     0.550                          0.583
## Residual Std. Error    1,802,737.000 (df = 37)        1,735,742.000 (df = 41)
## F Statistic             6.328*** (df = 11; 37)         10.570*** (df = 7; 41)
## =================================================================================
## Note:                                                 *p<0.1; **p<0.05; ***p<0.01
```

```r
knitr::kable(viewModelSummaryVIF(step_mdl))
```

var	Estimate	Std.Error	t-value	Pr(>|t|)	Significance	vif
chngdisc	2.149e+05	4.804e+04	4.474	5.98e-05	***	1.483179

var	Estimate	Std.Error	t-value	Pr(>|t|)	Significance	vif
chnglist	3.300e-04	1.527e-04	2.161	0.036598	*	1.153788
lagChngdisc	7.807e+04	4.607e+04	1.695	0.097747	.	1.364145
lagOnlineMar	4.337e-02	1.819e-02	2.384	0.021826	*	1.534020
n_saledays	3.276e+05	1.674e+05	1.957	0.057123	.	1.171176
Sponsorship	1.004e+05	5.176e+04	1.939	0.059384	.	1.355910
week	-7.162e+04	1.973e+04	-3.630	0.000779	***	1.356806

```r
pred_lm <- predict(step_mdl, model_data)
```

Regularized Linear Model:

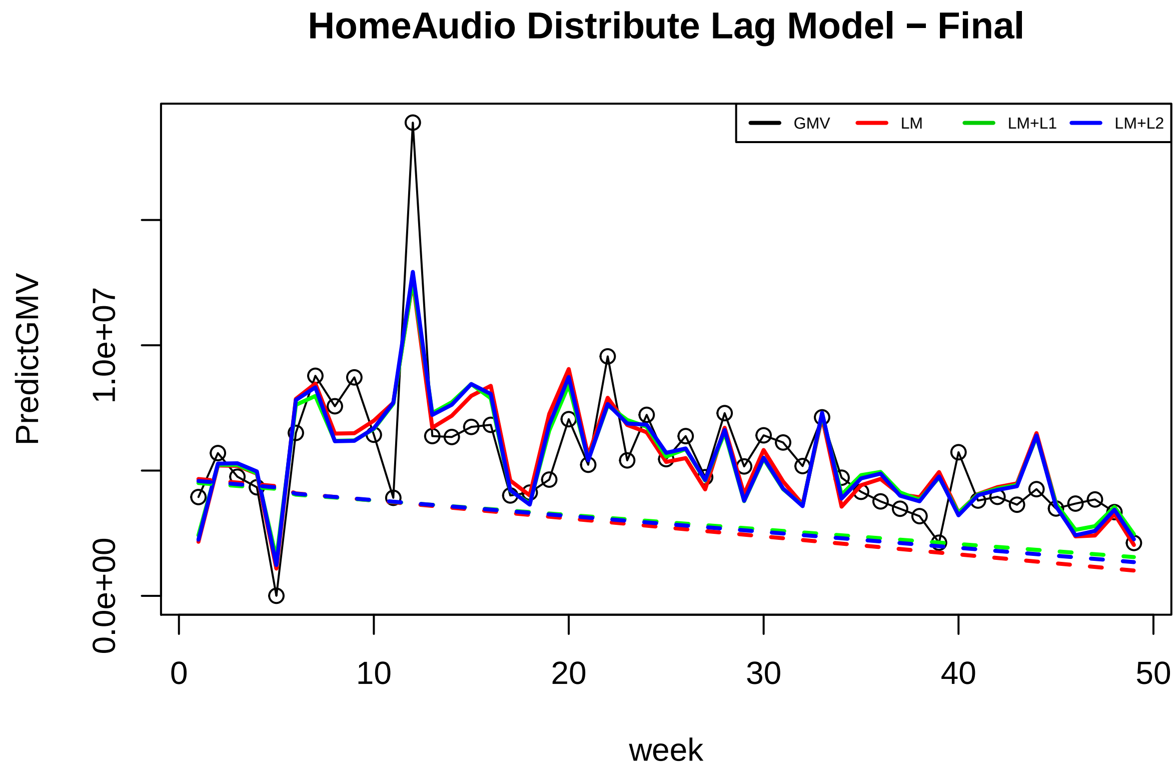
```r
x = as.matrix(subset(model_data, select=-gmv))
y = as.vector(model_data$gmv)

ridge_out <- atcLmReg(x,y,0,3)  # x, y, alpha, nfolds
lasso_out <- atcLmReg(x,y,1,3)  # x, y, alpha, nfolds
```

*

PLOTTING MODEL RESULTS

Plot Model prediction and base sales:

```
plot(model_data$gmv,main = 'HomeAudio Distribute Lag Model - Final',
     xlab='week',ylab='PredictGMV')
lines(model_data$gmv)
lines(pred_lm,col='red',lwd=2)
lines(ridge_out@pred,col='green',lwd=2)
lines(lasso_out@pred,col='blue',lwd=2)
lines(step_mdl$coefficients['(Intercept)']+step_mdl$coefficients['week']*model_data$week,
      lty=2,lwd=2,col='red')
lines(ridge_out@mdl$a0+ridge_out@mdl$beta['week',1]*model_data$week,
      lty=2,lwd=2,col='green')
lines(lasso_out@mdl$a0+lasso_out@mdl$beta['week',1]*model_data$week,
      lty=2,lwd=2,col='blue')
legend('topright',inset=0, legend=c('GMV','LM','LM+L1','LM+L2'),horiz = TRUE,
       lwd = 2, col=c(1:4), cex = 0.5)
```



*

*Model Coefficients:**

```r
coeff_lm <- as.data.frame(as.matrix(coef(step_mdl)))
coeff_l1 <- as.data.frame(as.matrix(coef(ridge_out@mdl)))
coeff_l2 <- as.data.frame(as.matrix(coef(lasso_out@mdl)))


lm_df=data.frame('x'=rownames(coeff_lm),'y'=coeff_lm)
colnames(lm_df) = c('coeff','lm')
l1_df=data.frame('x'=rownames(coeff_l1),'y'=coeff_l1)
colnames(l1_df)= c('coeff','l1')
l2_df=data.frame('x'=rownames(coeff_l2),'y'=coeff_l2)
colnames(l2_df) <- c('coeff','l2')

smry <- merge(lm_df,l1_df,all = TRUE)
smry <- merge(smry,l2_df,all=TRUE)

print(smry)
```

```
##             coeff           lm           l1           l2
## 1     (Intercept)  4.806894e+06  4.619322e+06  4.709296e+06
## 2        chngdisc  2.149101e+05  1.917799e+05  2.104381e+05
## 3        chnglist  3.299874e-04  2.881442e-04  3.187448e-04
## 4     lagChngdisc  7.806966e+04  6.896422e+04  8.018226e+04
## 5    lagOnlineMar  4.337476e-02  2.473527e-02  2.856749e-02
## 6        lagOther            NA  6.328879e-03  5.017867e-03
## 7          lagSEM            NA  1.718055e-02  1.653066e-02
## 8  lagSponsorship            NA  1.756964e+04  2.598458e+03
## 9      n_saledays  3.276245e+05  3.403996e+05  3.370734e+05
## 10          Other            NA  7.209192e-03  8.146978e-03
## 11    Sponsorship  1.003788e+05  9.160750e+04  9.832087e+04
## 12           week -7.162348e+04 -5.797088e+04 -6.343172e+04
```

```r
print(paste0('Ridge regression R2 : ',ridge_out@R2))
```

```
## [1] "Ridge regression R2 : 0.649983017356331"
```

```r
print(paste0('Lasso regression R2 : ',lasso_out@R2))
```

```
## [1] "Lasso regression R2 : 0.652893137143179"
```

```r
print(paste0('Linear Mode      R2 : ',getModelR2(step_mdl)))
```

```
## [1] "Multiple R-squared:  0.6435,\tAdjusted R-squared:  0.5826 "
## [1] "Linear Mode      R2 : Multiple R-squared:  0.6435,\tAdjusted R-squared:  0.5826 "
```

*

`Significant KPI`

Lasso(LM+L2) regression results a simple explainable model with significant KPIs as `Discount Inflation`, `Deliverycday`, `sale days`, `Sponsorship` `week,discount`,

```
# Model Optimization
```

```
# coeff            lm            l1            l2
# 1       (Intercept) -1.986887e+07  3.926035e+05 -7.858956e+06
# 2          chngdisc            NA  9.078256e+04  0.000000e+00
# 3          chnglist  3.417218e-04  1.514987e-04  3.087755e-04
# 4     deliverycdays            NA -1.013128e+05  0.000000e+00
# 5          discount  3.154408e+05  1.180970e+05  2.762660e+05
# 6       lagChngdisc            NA  2.808106e+04  8.658349e+03
# 7       lagChnglist            NA -3.442354e-05  1.085906e-05
# 8  lagdeliverycdays            NA -1.484086e+05 -8.604909e+04
# 9       lagdiscount            NA  8.398851e+03  0.000000e+00
# 10           laggmv -1.435463e-01 -8.573002e-02 -1.542386e-01
# 11      laglist_mrp            NA -9.081298e-05 -3.844650e-05
# 12           lagNPS  2.475900e-02  1.578393e-03  6.565710e-03
# 13     lagOnlineMar  1.204621e-01  2.987431e-03  4.550761e-03
# 14         lagOther            NA  4.101098e-03  8.788741e-03
# 15           lagSEM            NA  1.508486e-02  3.915611e-02
# 16   lagSponsorship            NA  5.201464e+04  3.335312e+03
# 17            lagTV            NA -2.444275e+05 -3.935209e+05
# 18         list_mrp            NA  1.507403e-04  0.000000e+00
# 19       n_saledays  2.882575e+05  2.312423e+05  2.764635e+05
# 20              NPS            NA -4.560996e-03  0.000000e+00
# 21  OnlineMarketing -1.163531e-01 -5.980353e-03 -3.060536e-02
# 22            Other  2.624659e-02  1.661755e-03  1.180933e-02
# 23              SEM            NA  6.218700e-03 -1.535383e-02
# 24      Sponsorship  3.447531e+05  8.413061e+04  2.499184e+05
# 25               TV -9.246923e+05 -2.773910e+04 -2.844342e+05
# 26             week -5.336887e+04 -2.033915e+04 -2.633297e+04
# [1] "Ridge regression R2 : 0.680677292912931"
# [1] "Lasso regression R2 : 0.72146688376956"
# [1] "Multiple R-squared:  0.7199,\tAdjusted R-squared:  0.6367 "
# [1] "Linear Mode      R2 :
#        Multiple R-squared:  0.7199,\tAdjusted R-squared:  0.6367 "
```

```
# coeff          lm            l1            l2
# 1     (Intercept)  4.806894e+06  4.626659e+06  4.709296e+06
# 2        chngdisc  2.149101e+05  1.933906e+05  2.104381e+05
# 3        chnglist  3.299874e-04  2.908324e-04  3.187448e-04
# 4     lagChngdisc  7.806966e+04  6.993309e+04  8.018226e+04
# 5    lagOnlineMar  4.337476e-02  2.501464e-02  2.856749e-02
# 6        lagOther            NA  6.280074e-03  5.017867e-03
# 7          lagSEM            NA  1.711051e-02  1.653066e-02
# 8  lagSponsorship            NA  1.665387e+04  2.598458e+03
# 9      n_saledays  3.276245e+05  3.404854e+05  3.370734e+05
# 10          Other            NA  7.279259e-03  8.146978e-03
# 11    Sponsorship  1.003788e+05  9.206706e+04  9.832087e+04
# 12           week -7.162348e+04 -5.844265e+04 -6.343172e+04
# [1] "Ridge regression R2 : 0.650428042810951"
# [1] "Lasso regression R2 : 0.652893137143179"
# [1] "Multiple R-squared:  0.6435,\tAdjusted R-squared:  0.5826 "
# [1] "Linear Mode      R2 :
#       Multiple R-squared:  0.6435,\tAdjusted R-squared:  0.5826 "
```