model_HA_LM_ad.R

atchirc

Tue May 23 01:26:18 2017

```r
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))
#

# # . . . . Ad Stock ----
model_data$adTV               <- as.numeric(
  stats::filter(model_data$TV,filter=0.5,method='recursive'))
model_data$adSponsorship      <- as.numeric(
  stats::filter(model_data$Sponsorship,filter=0.5,method='recursive'))
```

```r
model_data$adOnlineMarketing  <- as.numeric(
  stats::filter(model_data$OnlineMarketing,filter=0.5,method='recursive'))
model_data$adSEM              <- as.numeric(
  stats::filter(model_data$SEM,filter=0.5,method='recursive'))
model_data$adOther            <- as.numeric(
  stats::filter(model_data$Other,filter=0.5,method='recursive'))

model_data <- subset(model_data,select = -c(TV,Sponsorship,
                                            OnlineMarketing,
                                            SEM,Other))
```

*

```
**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

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

```
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.

```
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

```
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)
```

```r
  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(adTV,deliverycdays,NPS,
                                          chnglist,adOnlineMarketing,
                                          adOther,adSEM,discount,list_mrp))
```

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                             -33,198.650* (19,317.660)      -33,198.650* (19,317.660)
## n_saledays                      504,660.800*** (182,824.700)   504,660.800*** (182,824.700)
## chngdisc                        144,283.400*** (46,943.460)    144,283.400*** (46,943.460)
## adSponsorship                   122,580.400*** (30,065.780)    122,580.400*** (30,065.780)
## Constant                       4,085,217.000*** (795,710.500) 4,085,217.000*** (795,710.500)
## ----------------------------------------------------------------------------------------------
## Observations                                 50                             50
## R2                                          0.499                          0.499
## Adjusted R2                                 0.455                          0.455
## Residual Std. Error (df = 45)           2,040,508.000                  2,040,508.000
## F Statistic (df = 4; 45)                  11.218***                      11.218***
## ==============================================================================================
## 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
adSponsorship	122580	30066	4.077	0.000183	***	1.018979
chngdisc	144283	46944	3.074	0.003587	**	1.024878
n_saledays	504661	182825	2.760	0.008324	**	1.016151
week	-33199	19318	-1.719	0.092567	.	1.008470

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

Regularized Linear Model:

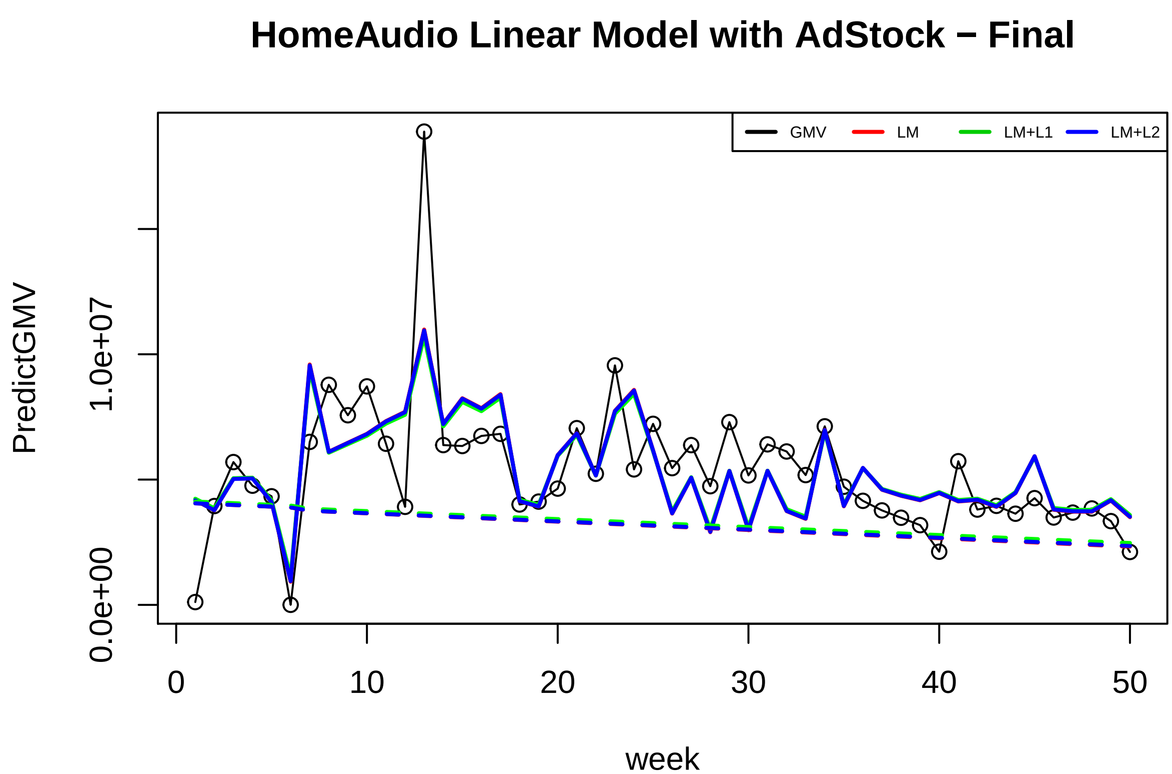
```r
x = as.matrix(subset(model_data, select=-gmv))
y = as.vector(model_data$gmv)

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$gmv,main = 'HomeAudio Linear Model with AdStock - Final',
     xlab='week',ylab='PredictGMV')
lines(model_data$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) 4085217.34 4153867.89 4084990.51
## 2 adSponsorship  122580.42  116403.15  121960.97
## 3      chngdisc  144283.40  137857.98  143331.58
## 4    n_saledays  504660.76  480945.58  500777.35
## 5          week  -33198.65  -31788.69  -32763.67
```

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

```
## [1] "Ridge regression R2 : 0.498152635209641"
```

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

```
## [1] "Lasso regression R2 : 0.499253055585335"
```

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

```
## [1] "Multiple R-squared:  0.4993,\tAdjusted R-squared:  0.4548 "
## [1] "Linear Mode      R2 : Multiple R-squared:  0.4993,\tAdjusted R-squared:  0.4548 "
```

*

Significant KPI

Lasso(LM+L1) regression results a simple explainable model with significant KPIs as `Discount Inflation`, `Deliverycday`, `sale days`, `Sponsorship Discount`, `week`, `NPS`

```
# Model Optimization
```

```
# coeff              lm            l1            l2
# 1         (Intercept) -1.334464e+07  4.123135e+06 -7.018625e+06
# 2  adOnlineMarketing            NA -7.181883e-03 -7.754690e-03
# 3            adOther            NA  5.943389e-03  1.254927e-02
# 4              adSEM            NA  8.895245e-03  3.724172e-03
# 5      adSponsorship  1.752083e+05  9.594346e+04  1.742541e+05
# 6               adTV -6.578450e+05 -3.261474e+05 -6.887309e+05
# 7           chngdisc            NA  8.554864e+04  5.620808e+04
# 8           chnglist  4.103389e-04  3.958495e-04  4.386169e-04
# 9      deliverycdays -2.522273e+05 -2.205743e+05 -2.584075e+05
# 10          discount  2.995054e+05  1.579444e+05  2.401065e+05
# 11          list_mrp -3.414011e-04 -2.660131e-04 -3.262682e-04
# 12        n_saledays  2.573460e+05  2.787913e+05  2.804350e+05
# 13               NPS  2.501028e-02  1.096844e-04  1.707140e-02
# 14              week            NA -1.212317e+04  0.000000e+00
# [1] "Ridge regression R2 : 0.646045984004396"
# [1] "Lasso regression R2 : 0.658554523295933"
# [1] "Multiple R-squared:  0.6392,\tAdjusted R-squared:  0.5688 "
# [1] "Linear Mode      R2 :
#         Multiple R-squared:  0.6392,\tAdjusted R-squared:  0.5688 "
```

```
# coeff          lm         l1         l2
# 1   (Intercept) 4085217.34 4153867.89 4084968.34
# 2 adSponsorship  122580.42  116403.15  121900.58
# 3      chngdisc  144283.40  137857.98  143238.78
# 4    n_saledays  504660.76  480945.58  500398.72
# 5          week  -33198.65  -31788.69  -32721.25
# [1] "Ridge regression R2 : 0.498152635209641"
# [1] "Lasso regression R2 : 0.49924806241621"
# [1] "Multiple R-squared:  0.4993,\tAdjusted R-squared:  0.4548 "
# [1] "Linear Mode      R2 :
#         Multiple R-squared:  0.4993,\tAdjusted R-squared:  0.4548 "
```