model_GA_DLag.R

atchirc

Sun May 21 23:53:15 2017

```
library(MASS)
library(car)
library(DataCombine)   # Pair wise correlation
library(stargazer)
library(dplyr)         # Data aggregation
library(glmnet)
source('./code/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=='CameraAccessory',
                     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))
#
# # . . . . NPS Inflation ----
# data$chngNPS  <- c(0,diff(data$NPS))
#
# # . . . . Lag List Price ----
# # Lag avg weekly list_mrp by 1 week
# data$lagListMrp <- data.table::shift(data$list_mrp)
```

```
#
# # . . . . Lag Discount ----
# # Lag weekly avg discount by 1 week
# model_data$lagDiscount <- data.table::shift(model_data$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)


#
# # . . . . Ad Stock ----
# data$adTotalInvestment  <- as.numeric(
#   stats::filter(data$TotalInvestment,filter=0.5,method='recursive'))
# data$adTV               <- as.numeric(
#   stats::filter(data$TV,filter=0.5,method='recursive'))
# data$adDigital          <- as.numeric(
#   stats::filter(data$Digital,filter=0.5,method='recursive'))
# data$adSponsorship      <- as.numeric(
#   stats::filter(data$Sponsorship,filter=0.5,method='recursive'))
# data$adContentMarketing <- as.numeric(
#   stats::filter(data$ContentMarketing,filter=0.5,method='recursive'))
# data$adOnlineMarketing  <- as.numeric(
#   stats::filter(data$OnlineMarketing,filter=0.5,method='recursive'))
# data$adAffiliates       <- as.numeric(
#   stats::filter(data$Affiliates,filter=0.5,method='recursive'))
# data$adSEM              <- as.numeric(
#   stats::filter(data$SEM,filter=0.5,method='recursive'))
# data$adRadio            <- as.numeric(
#   stats::filter(data$Radio,filter=0.5,method='recursive'))
# data$adOther            <- as.numeric(
#   stats::filter(data$Other,filter=0.5,method='recursive'))
# data$adNPS              <- as.numeric(
#   stats::filter(data$NPS,filter=0.5,method='recursive'))
```

*

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)
```

```
  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

```
# Prune KPI as part of model optimization
model_data <- na.omit(model_data)
model_data <- subset(model_data,select=-c(TV))
```

Linear Model:

```
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                      13,466.520 (58,178.980)
## discount                 328,115.600 (288,649.700)        118,772.300** (51,326.080)
## deliverycdays            489,974.800 (763,413.100)
## n_saledays               223,257.200 (182,813.500)        229,190.800 (142,859.700)
## Sponsorship              254,295.600 (211,900.500)       237,494.000*** (70,110.170)
## OnlineMarketing               0.026 (0.095)
## SEM                           -0.080 (0.062)                  -0.044** (0.021)
## Other                         0.005 (0.029)
## NPS                           -0.015 (0.044)
## list_mrp                      0.001 (0.0005)                0.0003** (0.0001)
## chnglist                     -0.0002 (0.0004)
## chngdisc                 -152,347.100 (223,171.800)
## laggmv                        -0.128 (0.178)
## lagdiscount
## lagdeliverycdays         -624,488.400 (822,356.700)
## lagTV                   -1,513,359.000 (998,965.500)
## lagSponsorship           104,523.600 (227,855.800)
## lagOnlineMar                  0.029 (0.111)                 0.038*** (0.013)
## lagSEM                        0.027 (0.066)
## lagOther                      0.029 (0.033)
## lagNPS                        0.024 (0.056)
## laglist_mrp
## lagChnglist                 0.0002 (0.0002)                0.0002** (0.0001)
## lagChngdisc              -41,629.340 (100,922.500)
## Constant                -23,089,460.000 (31,364,802.000) -5,850,186.000** (2,853,369.000)
## ------------------------------------------------------------------------------------------------
## Observations                        51                                   51
## R2                                0.687                                0.635
## Adjusted R2                       0.440                                0.576
```

```
## Residual Std. Error     1,776,163.000 (df = 28)          1,546,821.000 (df = 43)
## F Statistic              2.788*** (df = 22; 28)           10.686*** (df = 7; 43)
## ================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

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

var	Estimate	Std.Error	t-value	Pr(>|t|)	Significance	vif
discount	1.188e+05	5.133e+04	2.314	0.02551	*	1.486754
lagChnglist	2.332e-04	1.124e-04	2.075	0.04397	*	1.567389
lagOnlineMar	3.778e-02	1.327e-02	2.846	0.00675	**	1.249289
list_mrp	2.956e-04	1.105e-04	2.675	0.01053	*	1.477327
n_saledays	2.292e+05	1.429e+05	1.604	0.11597	NA	1.086798
SEM	-4.435e-02	2.116e-02	-2.096	0.04200	*	2.905998
Sponsorship	2.375e+05	7.011e+04	3.387	0.00152	**	3.341691

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

Regularized Linear Model:

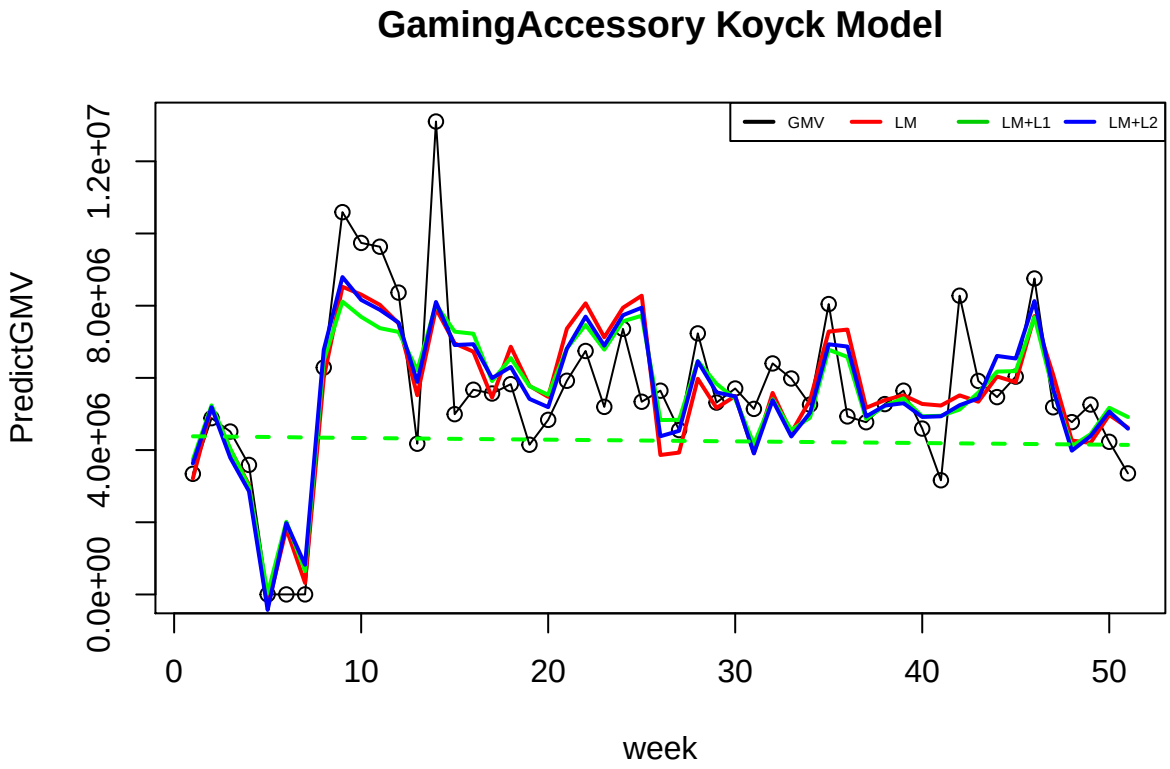
```
x = as.matrix(subset(model_data, select=-gmv))
y = as.vector(model_data$gmv)

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$gmv,main = 'GamingAccessory Koyck Model',
     xlab='week',ylab='PredictGMV')
lines(model_data$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:**

```
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) -5.850186e+06  4.390549e+06 -4.291621e+06
## 2          chngdisc            NA  2.449947e+04  0.000000e+00
## 3          chnglist            NA  1.160198e-04  5.691200e-05
## 4     deliverycdays            NA  4.355622e+04  0.000000e+00
## 5          discount  1.187723e+05  5.142912e+04  1.112255e+05
## 6       lagChngdisc            NA  8.996539e+02  0.000000e+00
## 7       lagChnglist  2.332116e-04  2.145849e-04  2.452673e-04
## 8  lagdeliverycdays            NA -4.239468e+03  0.000000e+00
## 9       lagdiscount            NA -1.275491e+04  0.000000e+00
## 10           laggmv            NA -6.264644e-03 -3.872416e-02
## 11      laglist_mrp            NA  1.675230e-05  0.000000e+00
## 12           lagNPS            NA -4.156010e-04  0.000000e+00
## 13     lagOnlineMar  3.778320e-02  1.291403e-02  7.562734e-03
## 14         lagOther            NA  4.138515e-03  1.370068e-02
## 15           lagSEM            NA -8.919481e-03  0.000000e+00
## 16   lagSponsorship            NA  7.306304e+04  9.269040e+04
## 17            lagTV            NA -2.273244e+05 -6.399569e+05
## 18         list_mrp  2.956362e-04  1.383065e-04  2.353227e-04
## 19       n_saledays  2.291908e+05  1.597431e+05  1.894902e+05
## 20              NPS            NA -8.221626e-03 -8.122346e-04
## 21  OnlineMarketing            NA  1.753697e-02  3.634779e-02
## 22            Other            NA  1.527987e-03  3.215995e-04
## 23              SEM -4.434695e-02 -1.511811e-02 -4.323863e-02
## 24      Sponsorship  2.374940e+05  1.033470e+05  1.845798e+05
## 25             week            NA -4.657119e+03  0.000000e+00
```

```
ridge_out@R2
```

```
## [1] 0.6358439
```

```
lasso_out@R2
```

```
## [1] 0.667608
```

*

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
```

```
# > print(smry)
# coeff            lm            l1            l2
# 1      (Intercept) -4.298317e+06  6.097679e+06  1.996419e+06
# 2         chngdisc            NA  1.706732e+04  1.532184e+04
# 3         chnglist            NA  1.508719e-05  0.000000e+00
# 4    deliverycdays            NA  1.706562e+05  5.297210e+04
# 5         discount  7.349317e+04  3.389389e+04  4.039341e+04
# 6      lagDiscount            NA -1.102003e+04  0.000000e+00
# 7           laggmv            NA -1.926869e-02 -1.596108e-02
# 8         list_mrp  3.394976e-04  2.541647e-04  2.868474e-04
# 9       n_saledays  2.476512e+05  2.305131e+05  2.325039e+05
# 10             NPS            NA -1.218913e-02 -7.241869e-03
# 11 OnlineMarketing  3.826100e-02  2.518933e-02  2.814779e-02
# 12           Other            NA  7.695146e-03  7.784236e-03
# 13             SEM -5.215457e-02 -3.559683e-02 -4.319949e-02
# 14     Sponsorship  2.577525e+05  2.059742e+05  2.542796e+05
# 15              TV            NA -2.060860e+05 -3.891445e+05
# 16            week            NA -1.598181e+04 -9.179892e+02
#
# > ridge_out@R2
# [1] 0.6122809
#
# > lasso_out@R2
# [1] 0.6156826
```

```
# > print(smry)
# coeff            lm            l1            l2
# 1      (Intercept) -4.141661e+05  7.713212e+06  4.878518e+06
# 2         chngdisc  3.675078e+04  3.718831e+04  5.076859e+04
# 3         chnglist            NA  3.375681e-05  7.272962e-06
# 4    deliverycdays            NA  1.970813e+05  3.347075e+05
# 5      lagDiscount            NA -2.050728e+03  2.902615e+04
# 6         list_mrp  2.891784e-04  2.280514e-04  2.605594e-04
# 7       n_saledays  2.364662e+05  2.297638e+05  2.769972e+05
# 8              NPS            NA -1.265991e-02 -1.060828e-02
# 9  OnlineMarketing  3.873164e-02  2.319657e-02  3.087915e-02
# 10           Other            NA  5.000504e-03  1.096254e-02
# 11             SEM -4.976103e-02 -3.297282e-02 -5.520338e-02
# 12     Sponsorship  2.616487e+05  1.881863e+05  2.535978e+05
# 13            week            NA -1.848834e+04 -4.003887e+04
#
# > ridge_out@R2
# [1] 0.6061716
#
```

```
# > lasso_out@R2
# [1] 0.6198029
```