model_GA_LM.R

atchirc

Sun May 21 22:38:36 2017

```r
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)
```

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

# # . . . . 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

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

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                      -27,388.840 (35,957.110)
## discount                  19,238.760 (116,525.700)        64,859.380 (47,885.430)
## deliverycdays             298,241.400 (272,022.700)
## n_saledays               281,516.900* (160,705.100)      249,425.100* (147,836.300)
## Sponsorship              265,547.100*** (81,150.720)    261,998.400*** (71,100.310)
## OnlineMarketing                0.042 (0.033)                  0.041*** (0.015)
## SEM                           -0.051* (0.026)                 -0.054** (0.021)
## Other                          0.011 (0.017)
## NPS                           -0.001 (0.020)
## list_mrp                      0.0003 (0.0002)                0.0004*** (0.0001)
## chnglist                    -0.00002 (0.0001)
## chngdisc                   27,769.430 (63,457.500)
## Constant                -1,086,548.000 (16,104,907.000) -4,205,266.000 (2,877,992.000)
## ------------------------------------------------------------------------------------
## Observations                        52                               52
## R2                                 0.644                            0.630
## Adjusted R2                        0.534                            0.581
## Residual Std. Error      1,689,116.000 (df = 39)        1,601,960.000 (df = 45)
## F Statistic              5.870*** (df = 12; 39)          12.778*** (df = 6; 45)
## ====================================================================================
## 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
discount	6.486e+04	4.789e+04	1.354	0.182349	NA	1.268696
list_mrp	3.520e-04	1.109e-04	3.176	0.002700	**	1.502641
n_saledays	2.494e+05	1.478e+05	1.687	0.098492	.	1.090761

var	Estimate	Std.Error	t-value	Pr(>|t|)	Significance	vif
OnlineMarketing	4.148e-02	1.460e-02	2.840	0.006747	**	1.460403
SEM	-5.363e-02	2.143e-02	-2.503	0.016022	*	2.801609
Sponsorship	2.620e+05	7.110e+04	3.685	0.000612	***	3.280275

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

Regularized Linear Model:

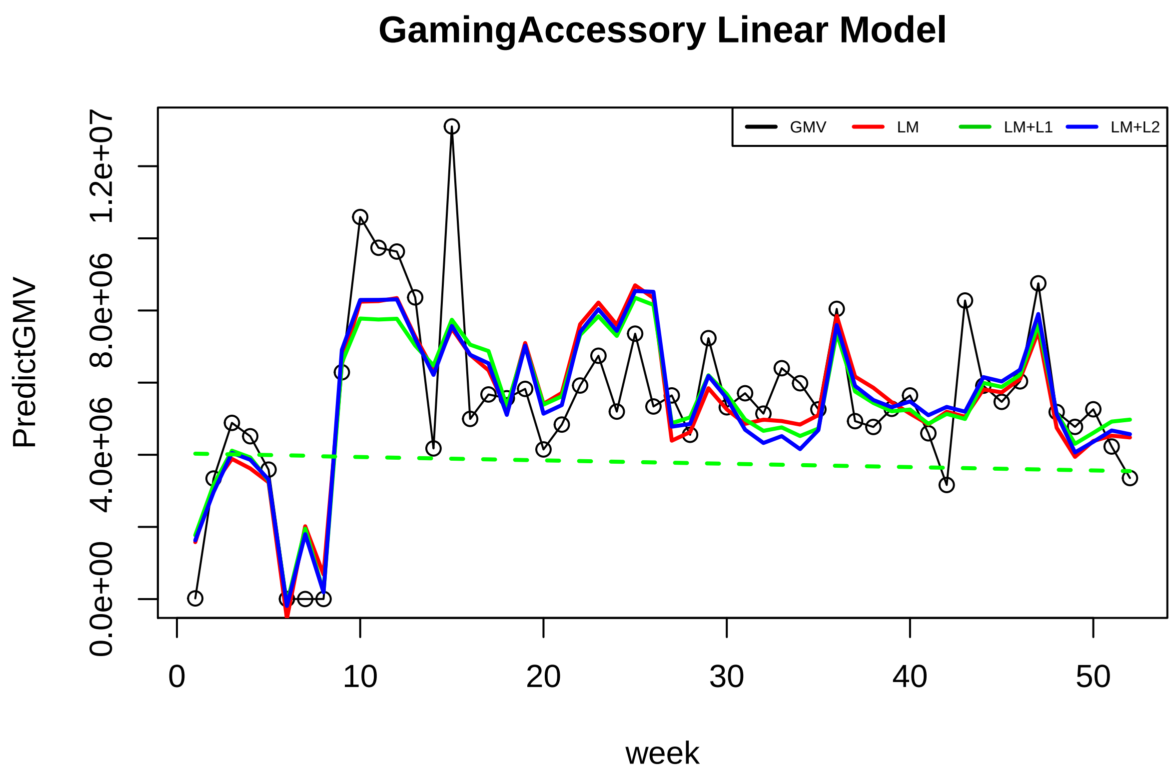
```
x = as.matrix(subset(model_data, select=-gmv))
y = as.vector(model_data$gmv)

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$gmv,main = 'GamingAccessory Linear Model',
     xlab='week',ylab='PredictGMV')
lines(model_data$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) -4.205266e+06  4.040125e+06 -8.207100e+05
## 2         chngdisc            NA  3.668112e+04  2.858341e+04
## 3         chnglist            NA  1.307862e-05 -1.368859e-05
## 4     deliverycdays           NA  1.674507e+05  2.868012e+05
## 5         discount  6.485938e+04  4.264987e+03  1.754609e+04
## 6         list_mrp  3.520229e-04  2.894513e-04  3.377355e-04
## 7       n_saledays  2.494251e+05  2.388797e+05  2.793114e+05
## 8              NPS            NA -8.325645e-03 -1.670934e-03
## 9  OnlineMarketing  4.147731e-02  2.798093e-02  4.132892e-02
## 10           Other            NA  5.322069e-03  1.049826e-02
## 11             SEM -5.362909e-02 -3.166888e-02 -5.049307e-02
## 12     Sponsorship  2.619984e+05  1.970039e+05  2.637522e+05
## 13            week            NA -9.303619e+03 -2.585423e+04
```

```
ridge_out@R2
```

```
## [1] 0.633043
```

```
lasso_out@R2
```

```
## [1] 0.6435987
```

*

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
```

```
# > print(smry)
# coeff              lm             l1             l2
# 1      (Intercept) -4.141661e+05  7.485367e+06  5.445941e+06
# 2         chngdisc  3.675078e+04  3.822982e+04  3.512998e+04
# 3         chnglist            NA  3.339202e-05  1.957373e-05
# 4     deliverycdays            NA  1.746820e+05  1.439615e+05
# 5      lagDiscount            NA  2.456224e+02  0.000000e+00
# 6         list_mrp  2.891784e-04  2.281347e-04  2.375811e-04
# 7       n_saledays  2.364662e+05  2.287571e+05  2.452622e+05
# 8              NPS            NA -1.243857e-02 -9.128712e-03
# 9  OnlineMarketing  3.873164e-02  2.444765e-02  2.941981e-02
# 10           Other            NA  6.323748e-03  8.512118e-03
# 11             SEM -4.976103e-02 -3.362561e-02 -4.682283e-02
# 12     Sponsorship  2.616487e+05  1.975294e+05  2.590272e+05
# 13              TV            NA -1.632189e+05 -3.544065e+05
# 14            week            NA -1.617192e+04 -1.343278e+04
#
# > ridge_out@R2
# [1] 0.6085013
#
# > lasso_out@R2
# [1] 0.6179322
```

```
# > print(smry)
# coeff              lm             l1             l2
# 1      (Intercept) -4.205266e+06  4.040125e+06 -8.028449e+05
# 2         chngdisc            NA  3.668112e+04  2.865168e+04
# 3         chnglist            NA  1.307862e-05 -1.342156e-05
# 4     deliverycdays            NA  1.674507e+05  2.858037e+05
# 5         discount  6.485938e+04  4.264987e+03  1.740505e+04
# 6         list_mrp  3.520229e-04  2.894513e-04  3.374212e-04
# 7       n_saledays  2.494251e+05  2.388797e+05  2.790864e+05
# 8              NPS            NA -8.325645e-03 -1.686547e-03
# 9  OnlineMarketing  4.147731e-02  2.798093e-02  4.127429e-02
# 10           Other            NA  5.322069e-03  1.045713e-02
# 11             SEM -5.362909e-02 -3.166888e-02 -5.042236e-02
# 12     Sponsorship  2.619984e+05  1.970039e+05  2.635711e+05
# 13            week            NA -9.303619e+03 -2.571906e+04
#
# > ridge_out@R2
# [1] 0.633043
#
# > lasso_out@R2
# [1] 0.6435931
```