model_GA_Kyock.R

atchirc

Mon May 22 16:27:15 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=='GamingAccessory',
                     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 GMV ----
# # Lag weekly avg discount by 1 week
model_data$laggmv <- data.table::shift(model_data$gmv)
```

*

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

```
# Prune KPI as part of model optimization
model_data <- na.omit(model_data)
model_data <- subset(model_data,select=-c(NPS,SEM,list_mrp,discount,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                      1,460.376 (20,747.430)
## deliverycdays           133,484.100 (159,638.700)      150,221.200* (80,097.860)
## n_saledays               95,913.430 (97,685.880)
## Sponsorship             80,058.790** (37,106.490)      84,099.980** (32,759.400)
## OnlineMarketing               0.020 (0.013)                0.026** (0.010)
## Other                         0.013 (0.010)                 0.014 (0.009)
## chnglist                     0.0001 (0.0001)
## chngdisc                43,127.320*** (15,480.100)     39,999.040*** (14,444.200)
## laggmv                        0.102 (0.139)
## Constant              1,240,929.000*** (429,121.000) 1,411,293.000*** (338,070.000)
## ----------------------------------------------------------------------------------
## Observations                       52                             52
## R2                               0.536                          0.507
## Adjusted R2                      0.436                          0.453
## Residual Std. Error    1,022,880.000 (df = 42)        1,007,418.000 (df = 46)
## F Statistic              5.384*** (df = 9; 42)          9.451*** (df = 5; 46)
## ==================================================================================
## 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
chngdisc	4.000e+04	1.444e+04	2.769	0.008075	**	1.023098
deliverycdays	1.502e+05	8.010e+04	1.875	0.067083	.	1.064676
OnlineMarketing	2.571e-02	1.032e-02	2.491	0.016390	*	1.855641
Other	1.420e-02	9.444e-03	1.504	0.139486	NA	1.551560
Sponsorship	8.410e+04	3.276e+04	2.567	0.013569	*	1.770872

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

Regularized Linear Model:

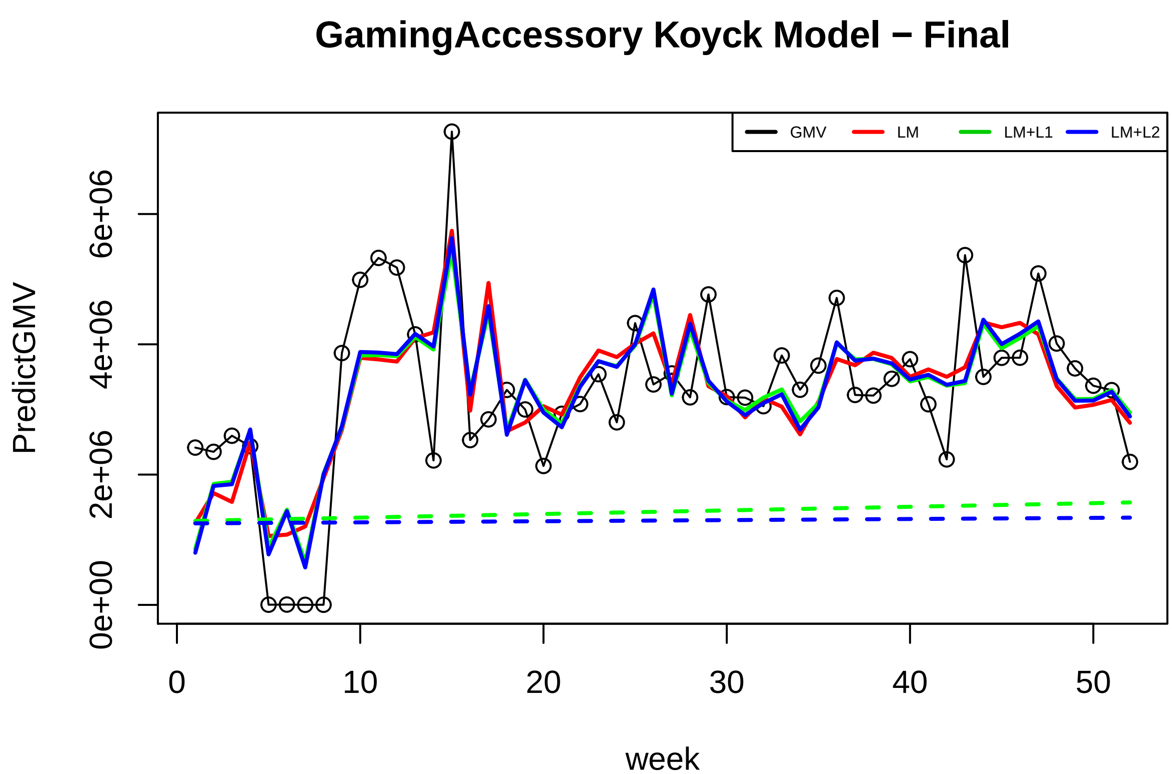
```r
x = as.matrix(subset(model_data, select=-gmv))
y = as.vector(model_data$gmv)

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$gmv,main = 'GamingAccessory Koyck Model - Final',
     xlab='week',ylab='PredictGMV')
lines(model_data$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('********koyck**********')
```

```
## [1] "********koyck**********"
```

```r
print(smry)
```

```
##              coeff           lm           l1           l2
## 1      (Intercept) 1.411293e+06 1.278884e+06 1.247807e+06
## 2        chngdisc 3.999904e+04 3.994247e+04 4.285293e+04
## 3        chnglist           NA 6.960671e-05 7.153673e-05
## 4    deliverycdays 1.502212e+05 9.214931e+04 1.297663e+05
## 5           laggmv           NA 1.072814e-01 1.016520e-01
## 6       n_saledays           NA 8.651000e+04 9.388715e+04
## 7  OnlineMarketing 2.571192e-02 1.895477e-02 2.003489e-02
## 8            Other 1.420094e-02 1.006338e-02 1.229718e-02
## 9      Sponsorship 8.409998e+04 7.312520e+04 7.942033e+04
## 10            week           NA 5.543696e+03 1.725794e+03
```

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

```
## [1] "Ridge regression R2 : 0.533480498452185"
```

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

```
## [1] "Lasso regression R2 : 0.535672498562392"
```

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

```
## [1] "Multiple R-squared:  0.5067,\tAdjusted R-squared:  0.4531 "
## [1] " Linear regression R2 : Multiple R-squared:  0.5067,\tAdjusted R-squared:  0.4531 "
```

*

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)  6.996400e+06  5.242092e+06  6.573439e+06
# 2         chngdisc            NA  2.067245e+04  9.468553e+03
# 3         chnglist            NA -5.859227e-06 -2.518083e-05
# 4    deliverycdays  2.800627e+05  1.267642e+05  2.237512e+05
# 5         discount  6.104641e+04  2.695146e+04  4.322371e+04
# 6           laggmv            NA -2.429706e-02 -9.547955e-02
# 7         list_mrp            NA  1.128403e-04  9.730563e-05
# 8       n_saledays            NA  9.596671e+04  1.106875e+05
# 9              NPS -1.582015e-02 -1.087928e-02 -1.451824e-02
# 10 OnlineMarketing            NA  9.345294e-03  5.220190e-03
# 11           Other            NA  5.628346e-03  7.942529e-03
# 12             SEM -4.453261e-02 -2.543729e-02 -4.248973e-02
# 13     Sponsorship  1.193868e+05  9.904806e+04  1.469937e+05
# 14              TV  5.205743e+05  1.351803e+05  1.852058e+05
# 15            week            NA  4.670993e+03  2.483487e+02
# > ridge_out@R2
# [1] 0.6338157
# > lasso_out@R2
# [1] 0.6522313
```

```
# coeff            lm           l1           l2
# 1      (Intercept) 1.411293e+06 1.270940e+06 1.247807e+06
# 2         chngdisc 3.999904e+04 4.044576e+04 4.285293e+04
# 3         chnglist           NA 7.013464e-05 7.153673e-05
# 4    deliverycdays 1.502212e+05 9.678533e+04 1.297663e+05
# 5           laggmv           NA 1.065741e-01 1.016520e-01
# 6       n_saledays           NA 8.783724e+04 9.388715e+04
# 7  OnlineMarketing 2.571192e-02 1.907708e-02 2.003489e-02
# 8            Other 1.420094e-02 1.038609e-02 1.229718e-02
# 9      Sponsorship 8.409998e+04 7.419512e+04 7.942033e+04
# 10            week           NA 5.135436e+03 1.725794e+03
# [1] "Ridge regression R2 : 0.534063573131462"
# [1] "Lasso regression R2 : 0.535672498562392"
# [1] "Multiple R-squared:  0.5067,\tAdjusted R-squared:  0.4531 "
# [1] " Linear regression R2 :
#        Multiple R-squared:  0.5067,\tAdjusted R-squared:  0.4531 "
# >
```