model_GA_DLag.R

atchirc

Mon May 22 16:29:29 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 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

```
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(lagdiscount,laggmv,lagTV,NPS,lagNPS,
                                          laglist_mrp,lagSEM,SEM,discount,TV,
                                          lagSponsorship))
```

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                   -6,243.352 (22,679.650)
## deliverycdays           625.362 (382,966.600)
## n_saledays             105,701.100 (98,101.200)
## Sponsorship            89,716.240** (36,791.860)     77,313.410** (30,064.100)
## OnlineMarketing            -0.008 (0.028)
## Other                       0.008 (0.014)
## list_mrp                  0.00003 (0.0002)
## chnglist                   0.0002 (0.0001)              0.0002** (0.0001)
## chngdisc              58,774.790*** (16,652.040)    59,723.120*** (15,350.800)
## lagdeliverycdays       160,288.300 (362,057.600)      102,642.800 (77,535.210)
## lagOnlineMar                0.029 (0.027)                0.024*** (0.009)
## lagOther                    0.009 (0.014)                 0.012 (0.009)
## lagChnglist                0.0002 (0.0001)              0.0002** (0.0001)
## lagChngdisc            35,914.920** (17,007.720)     35,610.840** (15,597.390)
## Constant             1,396,198.000 (1,540,601.000) 1,645,825.000*** (323,911.700)
## --------------------------------------------------------------------------------
## Observations                     52                             52
## R2                              0.615                         0.598
## Adjusted R2                     0.470                         0.523
## Residual Std. Error     991,980.500 (df = 37)         941,163.700 (df = 43)
## F Statistic            4.227*** (df = 14; 37)         7.981*** (df = 8; 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
chngdisc	5.972e+04	1.535e+04	3.891	0.000343	***	1.323978
chnglist	1.741e-04	7.487e-05	2.326	0.024801	*	1.373271
lagChngdisc	3.561e+04	1.560e+04	2.283	0.027426	*	1.364412
lagChnglist	1.554e-04	7.564e-05	2.054	0.046071	*	1.401844
lagdeliverycdays	1.026e+05	7.754e+04	1.324	0.192557	NA	1.091427
lagOnlineMar	2.379e-02	8.774e-03	2.712	0.009581	**	1.597205
lagOther	1.209e-02	9.049e-03	1.336	0.188546	NA	1.632176
Sponsorship	7.731e+04	3.006e+04	2.572	0.013663	*	1.708836

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

Regularized Linear Model:

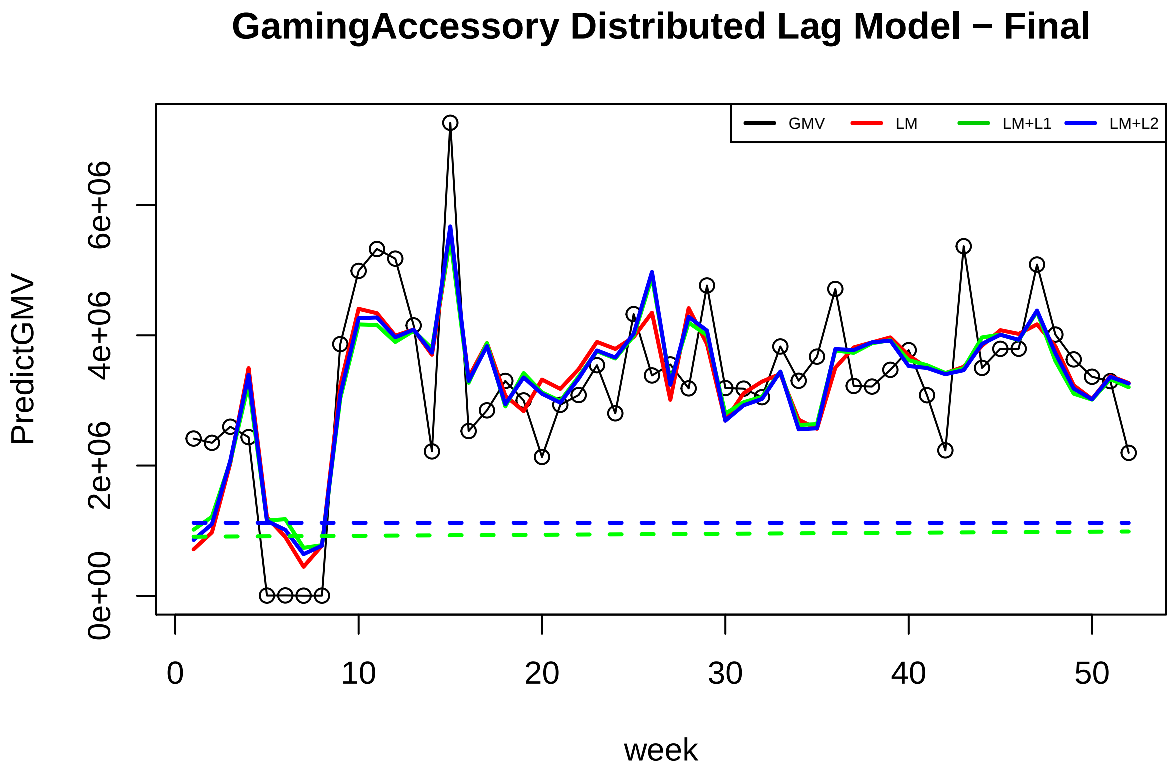
```
x = as.matrix(subset(model_data, select=-gmv))
y = as.vector(model_data$gmv)

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:

```r
plot(model_data$gmv,main = 'GamingAccessory Distributed Lag Model - Final',
     xlab='week',ylab='PredictGMV')
lines(model_data$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:**

```
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) 1.645825e+06 9.039253e+05 1.119355e+06
## 2          chngdisc 5.972312e+04 4.997126e+04 5.612873e+04
## 3          chnglist 1.741380e-04 9.537975e-05 1.289642e-04
## 4     deliverycdays           NA 2.919716e+04 0.000000e+00
## 5       lagChngdisc 3.561084e+04 2.763679e+04 3.305151e+04
## 6       lagChnglist 1.553776e-04 1.165036e-04 1.355250e-04
## 7  lagdeliverycdays 1.026428e+05 6.334367e+04 1.097015e+05
## 8      lagOnlineMar 2.379298e-02 1.468385e-02 2.121392e-02
## 9          lagOther 1.209038e-02 5.616374e-03 7.232751e-03
## 10         list_mrp           NA 8.163754e-05 5.497362e-05
## 11        n_saledays           NA 1.028100e+05 1.029384e+05
## 12  OnlineMarketing           NA 7.872448e-03 0.000000e+00
## 13            Other           NA 6.418893e-03 6.949971e-03
## 14      Sponsorship 7.731341e+04 7.549492e+04 8.486148e+04
## 15             week           NA 1.564481e+03 0.000000e+00
```

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

```
## [1] "Ridge regression R2 : 0.60736623168614"
```

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

```
## [1] "Lasso regression R2 : 0.613501532946416"
```

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

```
## [1] "Multiple R-squared:  0.5976,\tAdjusted R-squared:  0.5227 "
## [1] "Linear Mode      R2 : Multiple R-squared:  0.5976,\tAdjusted R-squared:  0.5227 "
```

*

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)  9.262345e+06  5.175952e+06  1.450808e+06
# 2          chngdisc  4.871132e+04  3.189364e+04  5.059131e+04
# 3          chnglist -2.200001e-04  4.328805e-05  0.000000e+00
# 4     deliverycdays  6.801050e+05  8.197115e+04  3.773325e+05
# 5          discount            NA  1.659154e+04  0.000000e+00
# 6       lagChngdisc  2.907198e+04  2.602593e+04  4.404848e+04
# 7       lagChnglist            NA  6.368811e-05 -8.760738e-05
# 8  lagdeliverycdays -6.222332e+05  3.078718e+03 -5.474381e+05
# 9       lagdiscount            NA -1.472181e+04 -2.009457e+04
# 10           laggmv            NA -5.452554e-02 -3.211260e-02
# 11      laglist_mrp            NA  2.963289e-05  3.688773e-04
# 12           lagNPS            NA  1.640285e-03  2.740288e-02
# 13     lagOnlineMar            NA  9.947901e-03  4.038448e-02
# 14         lagOther            NA  1.006683e-02  1.244483e-02
# 15           lagSEM  6.190645e-02  6.149363e-03  5.180175e-02
# 16   lagSponsorship -1.873398e+05  1.014581e+04 -8.088801e+04
# 17            lagTV -5.600746e+05 -5.683163e+05 -1.638005e+06
# 18         list_mrp  2.596167e-04  1.045144e-04  1.452596e-04
# 19       n_saledays  1.794812e+05  1.064671e+05  1.826192e+05
# 20              NPS -1.756014e-02 -9.978949e-03 -3.296579e-02
# 21  OnlineMarketing            NA  6.408966e-03 -1.356808e-02
# 22            Other  2.209605e-02  3.755186e-03  9.261901e-03
# 23              SEM -7.817009e-02 -2.146399e-02 -5.933397e-02
# 24      Sponsorship  3.131104e+05  9.238149e+04  2.013709e+05
# 25               TV            NA  3.588414e+05  7.983248e+05
# 26             week            NA  6.803727e+02  1.192395e+04
# > ridge_out@R2
# [1] 0.6785878
# > lasso_out@R2
# [1] 0.7620122
# lagdiscount,laggmv,lagTV,NPS,SEM
```

```
# coeff           lm          l1          l2
# 1       (Intercept) 1.645825e+06 9.039253e+05 1.121402e+06
# 2          chngdisc 5.972312e+04 4.997126e+04 5.626446e+04
# 3          chnglist 1.741380e-04 9.537975e-05 1.297062e-04
# 4     deliverycdays           NA 2.919716e+04 0.000000e+00
# 5       lagChngdisc 3.561084e+04 2.763679e+04 3.320145e+04
# 6       lagChnglist 1.553776e-04 1.165036e-04 1.361438e-04
# 7  lagdeliverycdays 1.026428e+05 6.334367e+04 1.104336e+05
# 8      lagOnlineMar 2.379298e-02 1.468385e-02 2.118047e-02
# 9          lagOther 1.209038e-02 5.616374e-03 7.305724e-03
# 10         list_mrp           NA 8.163754e-05 5.447224e-05
# 11       n_saledays           NA 1.028100e+05 1.034539e+05
# 12  OnlineMarketing           NA 7.872448e-03 0.000000e+00
# 13            Other           NA 6.418893e-03 6.958269e-03
# 14      Sponsorship 7.731341e+04 7.549492e+04 8.506770e+04
# 15             week           NA 1.564481e+03 0.000000e+00
# [1] "Ridge regression R2 : 0.60736623168614"
# [1] "Lasso regression R2 : 0.613565194881325"
# [1] "Multiple R-squared:  0.5976,\tAdjusted R-squared:  0.5227 "
# [1] "Linear Mode      R2 :
#          Multiple R-squared:  0.5976,\tAdjusted R-squared:  0.5227 "
```