# model_GA_Kyock_ad.R

*arman*

*Sat May 27 13:38:51 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))
#

# # . . . . 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'))
model_data$adOnlineMarketing  <- as.numeric(
```

```r
  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'))

# Prune regular
model_data <- subset(model_data,select = -c(TV,Sponsorship,
                                            OnlineMarketing,
                                            SEM,Other))

# # . . . . 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

---

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

**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                   -9,844.058 (24,905.950)
## deliverycdays          80,959.740 (172,366.800)
## n_saledays             72,670.260 (95,494.730)
## chnglist                  717.016 (704.947)
## chngdisc              35,199.780** (15,284.480)      35,170.650** (13,815.350)
## adTV                  -372,896.900 (238,656.900)    -387,193.000* (210,791.200)
## adSponsorship              0.013*** (0.004)               0.013*** (0.003)
## adOnlineMarketing          0.020** (0.008)                0.018*** (0.006)
## adSEM                     -0.029*** (0.010)              -0.028*** (0.008)
## adOther                    0.010 (0.007)                  0.009 (0.006)
## laggmv                    -0.037 (0.151)
## Constant            1,712,356.000*** (433,008.900) 1,610,585.000*** (314,059.400)
## ---------------------------------------------------------------------------------
## Observations                    52                             52
## R2                             0.581                          0.559
## Adjusted R2                    0.465                          0.500
## Residual Std. Error     996,028.100 (df = 40)          963,489.700 (df = 45)
## F Statistic            5.036*** (df = 11; 40)         9.492*** (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 |
|---|---|---|---|---|---|---|
| adOnlineMarketing | 1.848e-02 | 5.959e-03 | 3.101 | 0.003327 | ** | 2.548903 |
| adOther | 9.377e-03 | 6.365e-03 | 1.473 | 0.147683 | NA | 2.128693 |
| adSEM | -2.791e-02 | 8.113e-03 | -3.440 | 0.001264 | ** | 3.203060 |
| adSponsorship | 1.252e-02 | 2.990e-03 | 4.189 | 0.000129 | *** | 4.864565 |
| adTV | -3.872e+05 | 2.108e+05 | -1.837 | 0.072841 | . | 2.904520 |

| var | Estimate | Std.Error | t-value | Pr(>\|t\|) | Significance | vif |
|---|---|---|---|---|---|---|
| chngdisc | 3.517e+04 | 1.382e+04 | 2.546 | 0.014399 | * | 1.029574 |

```
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
```

*

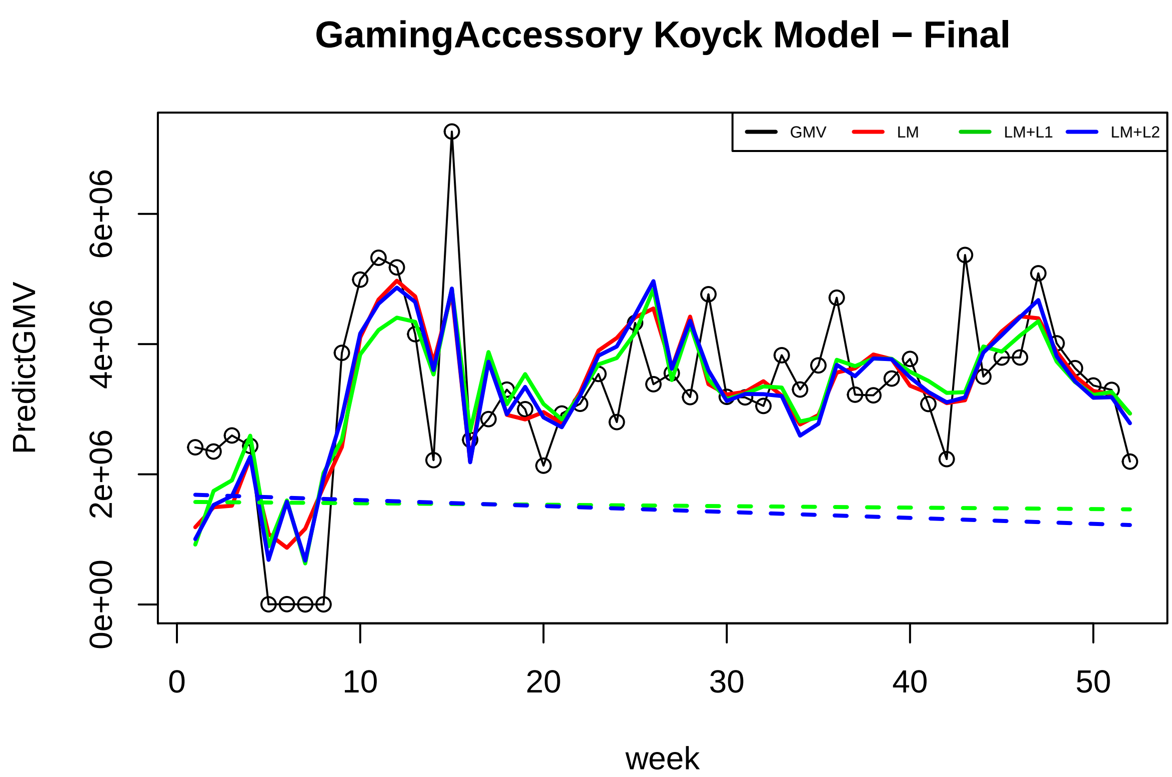
---

PLOTTING MODEL RESULTS

---

**Plot Model prediction and base sales:**

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

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

```
print(smry)
```

```
##                 coeff            lm            l1            l2
## 1         (Intercept)  1.610585e+06  1.579497e+06  1.703611e+06
## 2  adOnlineMarketing  1.847728e-02  1.405602e-02  1.948200e-02
## 3             adOther  9.377063e-03  6.505827e-03  1.017783e-02
## 4               adSEM -2.791111e-02 -1.614059e-02 -2.824369e-02
## 5       adSponsorship  1.252362e-02  7.963808e-03  1.255323e-02
## 6                adTV -3.871930e+05 -1.678232e+05 -3.659399e+05
## 7            chngdisc  3.517065e+04  3.745786e+04  3.526339e+04
## 8            chnglist            NA  7.907351e+02  7.168070e+02
## 9        deliverycdays            NA  5.538189e+04  7.653788e+04
## 10             laggmv            NA  6.517790e-02 -3.251217e-02
## 11          n_saledays            NA  6.816871e+04  7.214394e+04
## 12               week            NA -2.224492e+03 -9.117360e+03
```

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

```
## [1] "Ridge regression R2 : 0.557954744604513"
```

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

```
## [1] "Lasso regression R2 : 0.580682062591282"
```

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

```
## [1] "Multiple R-squared:  0.5586,\tAdjusted R-squared:  0.4998 "
## [1] " Linear regression R2 : Multiple R-squared:  0.5586,\tAdjusted R-squared:  0.4998 "
```

*

| Significant KPI |
|---|

```
#coeff           lm           l1           l2
#1        (Intercept)  1.610585e+06  1.579497e+06  1.704390e+06
#2  adOnlineMarketing  1.847728e-02  1.405602e-02  1.950633e-02
#3            adOther  9.377063e-03  6.505827e-03  1.019389e-02
#4              adSEM -2.791111e-02 -1.614059e-02 -2.827647e-02
#5      adSponsorship  1.252362e-02  7.963808e-03  1.256442e-02
#6               adTV -3.871930e+05 -1.678232e+05 -3.664399e+05
#7           chngdisc  3.517065e+04  3.745786e+04  3.525944e+04
#8           chnglist            NA  7.907351e+02  7.168335e+02
#9      deliverycdays            NA  5.538189e+04  7.696731e+04
#10            laggmv            NA  6.517790e-02 -3.284229e-02
#11        n_saledays            NA  6.816871e+04  7.219565e+04
#12              week            NA -2.224492e+03 -9.187108e+03

#[1] "Ridge regression R2 : 0.557954744604513"

#[1] "Lasso regression R2 : 0.580687778768009"

#[1] "Multiple R-squared:  0.5586,\tAdjusted R-squared:  0.4998 "
#[1] " Linear regression R2 : Multiple R-squared:  0.5586,\tAdjusted R-squared:  0.4998 "
```