

model_CA_DLag.R

atchirc

Mon May 22 22:23:55 2017

```
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 : Digital, ContentMarketing corr SEM
# delivery(b/c)days are corr, lets choose deliverydays
# 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))
#
# # . . . . Lag independant variables----
# # Lag weekly avg discount by 1 week
model_data$laggmvmv <- data.table::shift(model_data$gmvmv)
model_data$lagdiscout <- data.table::shift(model_data$discount)
model_data$lagdeliverydays <- data.table::shift(model_data$deliverydays)
```

```
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 optimization. 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,SEM,discount,lagTV,lagSEM,lagdiscount,
list_mrp,laglist_mrp,NPS,lagNPS))
```

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                -13,601.310 (40,162.640)
## deliverycdays       104,954.700 (644,458.300)
## n_saledays          184,265.800 (174,227.200)
## Sponsorship         131,316.400 (106,342.800)    143,030.300*** (43,962.240)
## OnlineMarketing      0.009 (0.054)
## Other                0.011 (0.026)
## chnglist             0.0003** (0.0001)          0.0003** (0.0001)
## chngdisc            116,975.100** (50,064.300)    76,492.920** (28,547.700)
## laggm               -0.0004 (0.157)
## lagdeliverycdays     17,084.190 (618,087.100)
## lagSponsorship       31,942.530 (103,784.200)
## lagOnlineMar         0.036 (0.058)              0.051*** (0.014)
## lagOther            -0.003 (0.027)
## lagChnglist          0.0004*** (0.0001)          0.0004*** (0.0001)
## lagChngdisc          52,080.050 (50,445.220)
## Constant            2,649,664.000*** (841,756.900) 2,554,898.000*** (547,561.200)
## -----
## Observations                51                51
## R2                          0.615                0.583
## Adjusted R2                  0.450                0.537
## Residual Std. Error    1,761,032.000 (df = 35)    1,615,445.000 (df = 45)
## F Statistic             3.726*** (df = 15; 35)    12.601*** (df = 5; 45)
## =====
## 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	7.649e+04	2.855e+04	2.679	0.010263	*	1.115905
chnglist	2.518e-04	9.702e-05	2.596	0.012702	*	1.080652
lagChnglist	3.548e-04	9.861e-05	3.598	0.000795	***	1.106676
lagOnlineMar	5.080e-02	1.359e-02	3.739	0.000520	***	1.200133
Sponsorship	1.430e+05	4.396e+04	3.253	0.002166	**	1.204647

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

Regularized Linear Model:

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

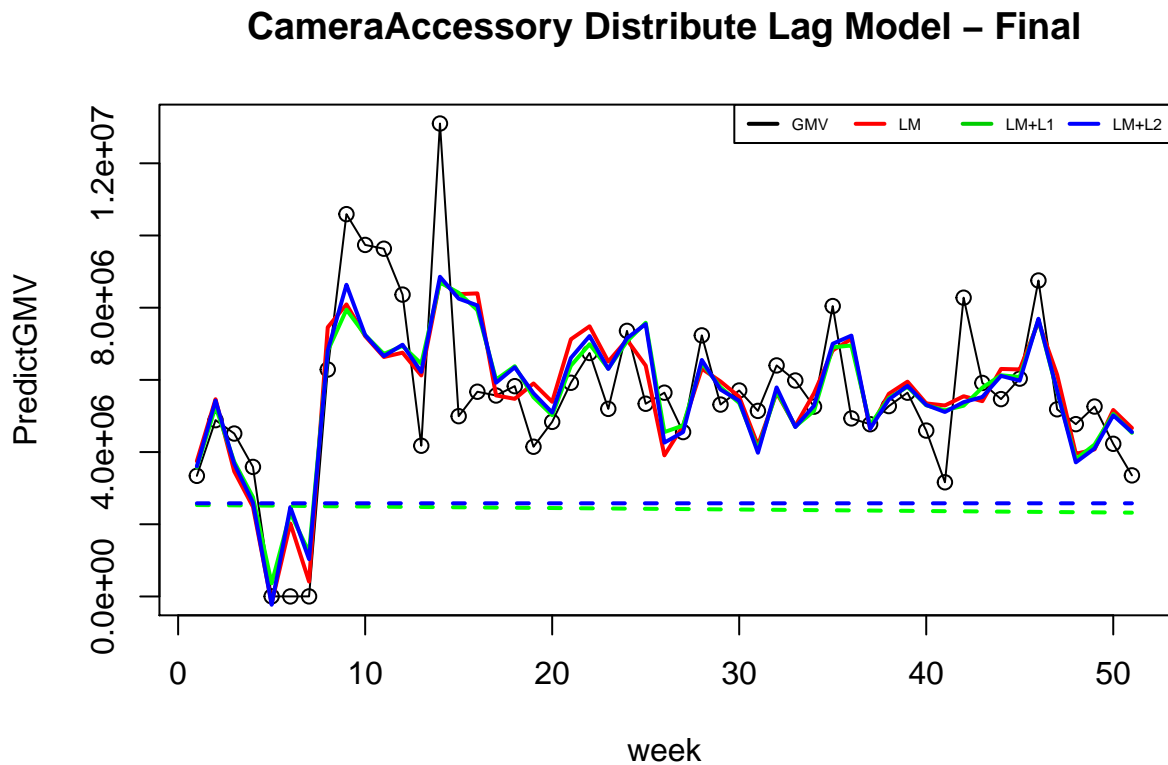
ridge_out <- atcLmReg(x,y,0,3) # x, y, alpha, nfolds
lasso_out <- atcLmReg(x,y,1,3) # x, y, alpha, nfolds
```

*

PLOTTING MODEL RESULTS

Plot Model prediction and base sales:

```
plot(model_data$gmvs, main = 'CameraAccessory Distribute Lag Model - Final',
     xlab='week', ylab='PredictGMV')
lines(model_data$gmvs)
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_mdls$coefficients['(Intercept)'] + step_mdls$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)	2.554898e+06	2.540746e+06	2.580690e+06	
## 2	chnghdisc	7.649292e+04	6.772848e+04	9.421654e+04	
## 3	chnghlist	2.518427e-04	2.154258e-04	2.518285e-04	
## 4	deliverycdays	NA	2.663430e+04	1.327867e+03	
## 5	lagChnghdisc	NA	9.238098e+03	3.022984e+04	
## 6	lagChnghlist	3.547880e-04	2.963856e-04	3.680278e-04	
## 7	lagdeliverycdays	NA	2.395941e+04	5.624987e+03	
## 8	laggmV	NA	5.339514e-02	1.178122e-02	
## 9	lagOnlineMar	5.080372e-02	1.860891e-02	2.716557e-02	
## 10	lagOther	NA	1.840534e-03	0.000000e+00	
## 11	lagSponsorship	NA	5.116042e+04	4.194978e+04	
## 12	n_saledays	NA	1.707805e+05	1.587098e+05	
## 13	OnlineMarketing	NA	2.113136e-02	1.460830e-02	
## 14	Other	NA	2.763553e-03	3.231619e-03	
## 15	Sponsorship	1.430303e+05	9.546267e+04	1.139776e+05	
## 16	week	NA	-4.128234e+03	0.000000e+00	

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

```
## [1] "Ridge regression R2 : 0.598282857749493"
```

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

```
## [1] "Lasso regression R2 : 0.609622854319688"
```

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

```
## [1] "Multiple R-squared: 0.5834,\tAdjusted R-squared: 0.5371 "
```

```
## [1] "Linear Mode R2 : Multiple R-squared: 0.5834,\tAdjusted R-squared: 0.5371 "
```


*

Significant KPI

Lasso(LM+L2) regression results a simple explainable model with significant KPIs as Discount Inflation, Deliverycday, sale days, Sponsorship week,discout,

Model Optimization

#	coeff	lm	l1	l2
# 1	(Intercept)	-5.969071e+06	4.579498e+06	9.150469e+05
# 2	chnghdisc	NA	2.403055e+04	1.404392e+04
# 3	chnghlist	NA	1.147520e-04	7.288558e-05
# 4	deliverycdays	NA	4.048955e+04	0.000000e+00
# 5	discount	1.236327e+05	4.963671e+04	7.598971e+04
# 6	lagChnghdisc	NA	6.245309e+01	0.000000e+00
# 7	lagChnghlist	2.293409e-04	2.098331e-04	2.287341e-04
# 8	lagdeliverycdays	NA	-3.025699e+03	0.000000e+00
# 9	lagdiscount	NA	-1.335381e+04	0.000000e+00
# 10	laggm	NA	-2.620557e-03	0.000000e+00
# 11	laglist_mrp	NA	1.674141e-05	0.000000e+00
# 12	lagNPS	NA	-5.625944e-04	0.000000e+00
# 13	lagOnlineMar	4.084108e-02	1.262188e-02	6.688259e-03
# 14	lagOther	NA	3.739609e-03	5.058915e-03
# 15	lagSEM	NA	-8.660429e-03	0.000000e+00
# 16	lagSponsorship	NA	7.039559e+04	4.807799e+04
# 17	lagTV	NA	-1.925906e+05	-2.177454e+05
# 18	list_mrp	2.884431e-04	1.373816e-04	1.833042e-04
# 19	n_saledays	2.427904e+05	1.568555e+05	1.714789e+05
# 20	NPS	NA	-8.196588e-03	-6.228913e-03
# 21	OnlineMarketing	NA	1.710518e-02	2.425264e-02
# 22	Other	2.019568e-02	1.489752e-03	0.000000e+00
# 23	SEM	-5.086349e-02	-1.401646e-02	-3.032600e-02
# 24	Sponsorship	3.140919e+05	1.003712e+05	1.560263e+05
# 25	TV	-8.880872e+05	-1.956665e+04	0.000000e+00
# 26	week	NA	-4.224726e+03	0.000000e+00
# [1]	"Ridge regression R2 : 0.632501417802671"			
# [1]	"Lasso regression R2 : 0.645565137277216"			
# [1]	"Multiple R-squared: 0.6579, \tAdjusted R-squared: 0.5828 "			
# [1]	"Linear Mode R2 :			
#	Multiple R-squared: 0.6579, \tAdjusted R-squared: 0.5828 "			

#	coeff	lm	l1	l2
# 1	(Intercept)	2.554898e+06	4.913420e+06	5.846966e+06
# 2	chnghdisc	7.649292e+04	5.274919e+04	7.169863e+04
# 3	chnghlist	2.518427e-04	1.160990e-04	1.331778e-04
# 4	deliverycdays	NA	5.896428e+04	4.689963e+04
# 5	lagChnghdisc	NA	3.785165e+03	1.515076e+04
# 6	lagChnghlist	3.547880e-04	1.967177e-04	2.573714e-04
# 7	lagdeliverycdays	NA	2.027915e+04	0.000000e+00
# 8	laggm	NA	1.028940e-02	0.000000e+00
# 9	laglist_mrp	NA	1.873679e-05	0.000000e+00
# 10	lagNPS	NA	1.161537e-03	0.000000e+00

```

# 11      lagOnlineMar 5.080372e-02  1.027355e-02  1.248910e-02
# 12      lagOther      NA  2.953227e-03  0.000000e+00
# 13      lagSponsorship      NA  4.405822e+04  2.549686e+04
# 14      list_mrp      NA  1.399100e-04  1.249556e-04
# 15      n_saledays      NA  1.579620e+05  1.577822e+05
# 16      NPS      NA -7.935344e-03 -7.801763e-03
# 17      OnlineMarketing      NA  1.516293e-02  1.328937e-02
# 18      Other      NA  1.958178e-03  2.287910e-03
# 19      Sponsorship 1.430303e+05  8.229495e+04  1.044978e+05
# 20      week      NA -1.254369e+03  0.000000e+00
# [1] "Ridge regression R2 : 0.601140151803534"
# [1] "Lasso regression R2 : 0.611020467029655"
# [1] "Multiple R-squared:  0.5834,\tAdjusted R-squared:  0.5371 "
# [1] "Linear Mode      R2 :
#      Multiple R-squared:  0.5834,\tAdjusted R-squared:  0.5371 "

```