2-Models.R

arman

Sat May 20 22:29:21 2017

```
# ********************************************************************************
#                    LOAD LIBRARY ----
# ********************************************************************************
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.3.3
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(car)
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(Hmisc)   # describe
```

```
## Loading required package: lattice
```

```
## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     combine, src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
# ************************************************************************
#                   LOAD DATA ---- Product Categories ----
# ************************************************************************
camera_accessory_data <- read.csv('./intrim/cameraAccessory.csv')
home_audio_data <- read.csv('./intrim/homeAudio.csv')
gaming_accessory_data <- read.csv('./intrim/gamingAccessory.csv')

# ************************************************************************
#                   Create Training & Test Datasets ----
# ************************************************************************
#Lets divide the Train & Test data. For this we will use the first 36 weeks data as Train set and rest

#Camera Accessory
cam_train <- subset(camera_accessory_data, week <= 36)
cam_test <- subset(camera_accessory_data, week > 36)
cam_train <- cam_train[,-1]
cam_test <- cam_test[,-1]

#Gaming Accessory
gam_train <- subset(gaming_accessory_data, week <= 36)
gam_test <- subset(gaming_accessory_data, week > 36)
gam_train <- gam_train[,-1]
gam_test <- gam_test[,-1]

#Gaming Accessory
hom_train <- subset(home_audio_data, week <= 36)
hom_test <- subset(home_audio_data, week > 36)
hom_train <- hom_train[,-1]
hom_test <- hom_test[,-1]

# ************************************************************************
#                   MODELLING ---- Simple Linear Model ----
# ************************************************************************
```

Initial Linear Model

```
slm_cam1 <- lm(gmv~ .,data=cam_train)
```

Auto-Optimize Model

```
step_slm_cam <- stepAIC(slm_cam1, direction = "both",trace=FALSE)
summary(step_slm_cam)
```

```
##
## Call:
## lm(formula = gmv ~ product_mrp + list_price + Promotion + sla +
##     cat_mid + cat_premium + ContentMarketing + OnlineMarketing +
##     Affiliates + SEM + NPS, data = cam_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3801831  -310366   -44240   339634  5198354
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.004e+07  5.299e+06   1.895 0.061430 .
## product_mrp      -9.425e+01  2.270e+01  -4.152 7.79e-05 ***
## list_price        1.028e+02  3.239e+01   3.174 0.002095 **
## Promotion         1.015e+07  2.918e+06   3.478 0.000799 ***
## sla               1.769e+05  9.710e+04   1.821 0.072060 .
## cat_mid          -1.040e+06  7.079e+05  -1.470 0.145370
## cat_premium      -1.055e+06  6.386e+05  -1.651 0.102332
## ContentMarketing -1.244e+06  3.339e+05  -3.725 0.000351 ***
## OnlineMarketing  -1.296e+06  8.535e+05  -1.519 0.132590
## Affiliates        5.761e+06  3.179e+06   1.812 0.073458 .
## SEM               1.273e+05  8.375e+04   1.520 0.132241
## NPS              -2.244e+05  8.675e+04  -2.587 0.011396 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1123000 on 85 degrees of freedom
## Multiple R-squared:  0.8293, Adjusted R-squared:  0.8072
## F-statistic: 37.55 on 11 and 85 DF,  p-value: < 2.2e-16
```

```
#Pruning of Variables to arrive at Final Model
slm_cam2 <- lm(formula = gmv ~ product_mrp + list_price + Promotion + sla +
                 cat_mid + cat_premium + ContentMarketing + OnlineMarketing +
                 Affiliates + SEM + NPS, data = cam_train)
summary(slm_cam2)
```

```
##
## Call:
## lm(formula = gmv ~ product_mrp + list_price + Promotion + sla +
##     cat_mid + cat_premium + ContentMarketing + OnlineMarketing +
##     Affiliates + SEM + NPS, data = cam_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3801831  -310366   -44240   339634  5198354
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.004e+07  5.299e+06   1.895 0.061430 .
## product_mrp      -9.425e+01  2.270e+01  -4.152 7.79e-05 ***
## list_price        1.028e+02  3.239e+01   3.174 0.002095 **
## Promotion         1.015e+07  2.918e+06   3.478 0.000799 ***
## sla               1.769e+05  9.710e+04   1.821 0.072060 .
## cat_mid          -1.040e+06  7.079e+05  -1.470 0.145370
```

```
## cat_premium      -1.055e+06  6.386e+05  -1.651 0.102332
## ContentMarketing -1.244e+06  3.339e+05  -3.725 0.000351 ***
## OnlineMarketing  -1.296e+06  8.535e+05  -1.519 0.132590
## Affiliates        5.761e+06  3.179e+06   1.812 0.073458 .
## SEM               1.273e+05  8.375e+04   1.520 0.132241
## NPS              -2.244e+05  8.675e+04  -2.587 0.011396 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1123000 on 85 degrees of freedom
## Multiple R-squared:  0.8293, Adjusted R-squared:  0.8072
## F-statistic: 37.55 on 11 and 85 DF,  p-value: < 2.2e-16
```

```
vif(slm_cam2)
```

```
##      product_mrp       list_price        Promotion              sla
##        70.938298        75.718661        15.614113         1.410193
##          cat_mid      cat_premium ContentMarketing  OnlineMarketing
##         8.388153         6.578784         8.625826       262.080090
##       Affiliates              SEM              NPS
##       269.942274         6.071132         8.638058
```

```
#Removing Affiliates
slm_cam3 <- lm(formula = gmv ~ product_mrp + list_price + Promotion + sla +
                 cat_mid + cat_premium + ContentMarketing + OnlineMarketing +
                 SEM + NPS, data = cam_train)
summary(slm_cam3)
```

```
##
## Call:
## lm(formula = gmv ~ product_mrp + list_price + Promotion + sla +
##     cat_mid + cat_premium + ContentMarketing + OnlineMarketing +
##     SEM + NPS, data = cam_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4015730  -433140   -91221   403003  5256995
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.060e+07  5.360e+06   1.978 0.051181 .
## product_mrp      -9.296e+01  2.299e+01  -4.044 0.000114 ***
## list_price        1.017e+02  3.282e+01   3.100 0.002617 **
## Promotion         1.055e+07  2.948e+06   3.579 0.000570 ***
## sla               1.684e+05  9.827e+04   1.713 0.090279 .
## cat_mid          -9.228e+05  7.142e+05  -1.292 0.199827
## cat_premium      -1.036e+06  6.469e+05  -1.602 0.112878
## ContentMarketing -9.018e+05  2.792e+05  -3.230 0.001752 **
## OnlineMarketing   2.080e+05  2.021e+05   1.029 0.306372
## SEM               1.994e+03  4.789e+04   0.042 0.966895
## NPS              -2.282e+05  8.787e+04  -2.597 0.011062 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1137000 on 86 degrees of freedom
```

```
## Multiple R-squared:  0.8227, Adjusted R-squared:  0.8021
## F-statistic: 39.91 on 10 and 86 DF,  p-value: < 2.2e-16
```

```
vif(slm_cam3)
```

```
##      product_mrp       list_price        Promotion              sla
##        70.868670        75.692498        15.524103         1.406895
##          cat_mid      cat_premium ContentMarketing  OnlineMarketing
##         8.317740         6.577124         5.872703        14.314583
##              SEM              NPS
##         1.933951         8.632988
```

```
#Removing List_price
slm_cam4 <- lm(formula = gmv ~ product_mrp + Promotion + sla +
                 cat_mid + cat_premium + ContentMarketing + OnlineMarketing +
                 SEM + NPS, data = cam_train)
summary(slm_cam4)
```

```
##
## Call:
## lm(formula = gmv ~ product_mrp + Promotion + sla + cat_mid +
##     cat_premium + ContentMarketing + OnlineMarketing + SEM +
##     NPS, data = cam_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3053468  -521447   -60198   585272  5797805
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.575e+07  5.343e+06   2.947  0.00412 **
## product_mrp      -2.329e+01  5.050e+00  -4.612 1.36e-05 ***
## Promotion         2.129e+06  1.200e+06   1.774  0.07953 .
## sla               1.824e+05  1.029e+05   1.772  0.07984 .
## cat_mid          -2.667e+06  4.610e+05  -5.786 1.12e-07 ***
## cat_premium      -2.343e+06  5.145e+05  -4.553 1.71e-05 ***
## ContentMarketing -9.189e+05  2.926e+05  -3.141  0.00230 **
## OnlineMarketing   2.247e+05  2.118e+05   1.061  0.29162
## SEM              -5.390e+03  5.014e+04  -0.107  0.91464
## NPS              -2.583e+05  9.155e+04  -2.822  0.00592 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1192000 on 87 degrees of freedom
## Multiple R-squared:  0.8029, Adjusted R-squared:  0.7825
## F-statistic: 39.38 on 9 and 87 DF,  p-value: < 2.2e-16
```

```
vif(slm_cam4)
```

```
##      product_mrp        Promotion              sla          cat_mid
##         3.112501         2.340143         1.403910         3.153418
##      cat_premium ContentMarketing  OnlineMarketing              SEM
##         3.785604         5.870409        14.304349         1.929166
##              NPS
##         8.527313
```

```r
#Removing OnlineMarketing
slm_cam5 <- lm(formula = gmv ~ product_mrp + Promotion + sla +
                cat_mid + cat_premium + ContentMarketing +
                SEM + NPS, data = cam_train)
summary(slm_cam5)
```

```
##
## Call:
## lm(formula = gmv ~ product_mrp + Promotion + sla + cat_mid +
##     cat_premium + ContentMarketing + SEM + NPS, data = cam_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3390263  -530663   -52142   507915  5928818
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.059e+07  2.775e+06   7.420 6.96e-11 ***
## product_mrp      -2.254e+01  5.004e+00  -4.505 2.03e-05 ***
## Promotion         2.162e+06  1.200e+06   1.801  0.07510 .
## sla               1.692e+05  1.022e+05   1.655  0.10139
## cat_mid          -2.703e+06  4.601e+05  -5.875 7.38e-08 ***
## cat_premium      -2.417e+06  5.101e+05  -4.738 8.23e-06 ***
## ContentMarketing -6.835e+05  1.909e+05  -3.581  0.00056 ***
## SEM              -2.754e+04  4.563e+04  -0.604  0.54771
## NPS              -3.396e+05  5.014e+04  -6.773 1.36e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1193000 on 88 degrees of freedom
## Multiple R-squared:  0.8004, Adjusted R-squared:  0.7822
## F-statistic:  44.1 on 8 and 88 DF,  p-value: < 2.2e-16
```

```r
vif(slm_cam5)
```

```
##      product_mrp        Promotion              sla          cat_mid
##         3.051196         2.338553         1.383561         3.136402
##      cat_premium ContentMarketing              SEM              NPS
##         3.715657         2.494427         1.594906         2.554456
```

```r
#Removing ContentMarketing
slm_cam6 <- lm(formula = gmv ~ product_mrp + Promotion + sla +
                cat_mid + cat_premium +
                SEM + NPS, data = cam_train)
summary(slm_cam6)
```

```
##
## Call:
## lm(formula = gmv ~ product_mrp + Promotion + sla + cat_mid +
##     cat_premium + SEM + NPS, data = cam_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3626304  -541627  -196685   315840  6289418
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.515e+07  2.472e+06   6.130 2.36e-08 ***
## product_mrp -2.375e+01  5.314e+00  -4.469 2.31e-05 ***
## Promotion    2.158e+06  1.278e+06   1.689   0.0947 .
## sla          2.960e+04  1.006e+05   0.294   0.7693
## cat_mid     -2.745e+06  4.896e+05  -5.607 2.29e-07 ***
## cat_premium -2.378e+06  5.428e+05  -4.382 3.21e-05 ***
## SEM         -6.292e+04  4.741e+04  -1.327   0.1879
## NPS         -2.288e+05  4.199e+04  -5.449 4.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1270000 on 89 degrees of freedom
## Multiple R-squared:  0.7713, Adjusted R-squared:  0.7533
## F-statistic: 42.87 on 7 and 89 DF,  p-value: < 2.2e-16
```

```
vif(slm_cam6)
```

```
## product_mrp   Promotion         sla     cat_mid cat_premium         SEM
##    3.037384    2.338551    1.182243    3.134409    3.714010    1.520118
##         NPS
##    1.581360
```

```
#Removing Promotion
slm_cam7 <- lm(formula = gmv ~ product_mrp + sla +
                cat_mid + cat_premium +
                SEM + NPS, data = cam_train)
summary(slm_cam7)
```

```
##
## Call:
## lm(formula = gmv ~ product_mrp + sla + cat_mid + cat_premium +
##     SEM + NPS, data = cam_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3741899  -497812  -192033   285401  6444280
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.695e+07  2.253e+06   7.523 3.87e-11 ***
## product_mrp -2.564e+01  5.247e+00  -4.886 4.44e-06 ***
## sla         -1.654e+04  9.779e+04  -0.169    0.866
## cat_mid     -3.300e+06  3.665e+05  -9.003 3.39e-14 ***
## cat_premium -2.485e+06  5.447e+05  -4.562 1.60e-05 ***
## SEM         -6.660e+04  4.785e+04  -1.392    0.167
## NPS         -2.394e+05  4.194e+04  -5.707 1.45e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1283000 on 90 degrees of freedom
## Multiple R-squared:  0.7639, Adjusted R-squared:  0.7482
## F-statistic: 48.54 on 6 and 90 DF,  p-value: < 2.2e-16
```

```r
vif(slm_cam7)
```

```
## product_mrp         sla     cat_mid cat_premium         SEM         NPS
##    2.902105    1.095068    1.721575    3.664177    1.516903    1.546035
```

```r
#Removing SLA & SEM, this is our final model
slm_cam8 <- lm(formula = gmv ~ product_mrp +
                 cat_mid + cat_premium +
                 NPS, data = cam_train)
summary(slm_cam8)
```

```
##
## Call:
## lm(formula = gmv ~ product_mrp + cat_mid + cat_premium + NPS,
##     data = cam_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3630169  -476495  -165450   358933  6063081
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.493e+07  1.703e+06   8.770 8.82e-14 ***
## product_mrp -2.500e+01  5.165e+00  -4.840 5.21e-06 ***
## cat_mid     -3.296e+06  3.517e+05  -9.372 4.79e-15 ***
## cat_premium -2.511e+06  5.335e+05  -4.706 8.88e-06 ***
## NPS         -2.060e+05  3.416e+04  -6.033 3.34e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1283000 on 92 degrees of freedom
## Multiple R-squared:  0.7588, Adjusted R-squared:  0.7483
## F-statistic: 72.34 on 4 and 92 DF,  p-value: < 2.2e-16
```

```r
vif(slm_cam8)
```

```
## product_mrp     cat_mid cat_premium         NPS
##    2.812426    1.585370    3.516587    1.025489
```

```r
#Test the model on Test Dataset
pred_cam_slm<-predict(slm_cam8,cam_test[,-1])

#Add New column for predicted_gmv
cam_test$predicted_gmv <- pred_cam_slm

#Lets look at the Corr & R2
cor(cam_test$gmv, cam_test$predicted_gmv)
```

```
## [1] 0.9379132
```

```r
cor(cam_test$gmv, cam_test$predicted_gmv)^2
```

```
## [1] 0.8796811
```

```r
#xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
```

Initial Linear Model

```r
slm_gam1 <- lm(gmv~ .,data=gam_train)
```

Auto-Optimize Model

```r
step_slm_gam <- stepAIC(slm_gam1, direction = "both",trace=FALSE)
summary(step_slm_gam)
```

```
##
## Call:
## lm(formula = gmv ~ Promotion + cat_mid + sale_days + Sponsorship +
##     NPS, data = gam_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1544179  -440047   -44635   343456  2379851
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21240855    2626057   8.088 2.78e-11 ***
## Promotion    5622738    2314864   2.429   0.0181 *
## cat_mid     -6284504    1306841  -4.809 1.01e-05 ***
## sale_days      96290      55152   1.746   0.0858 .
## Sponsorship   -39262       7690  -5.106 3.37e-06 ***
## NPS          -382603      49635  -7.708 1.27e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 760900 on 62 degrees of freedom
## Multiple R-squared:  0.8375, Adjusted R-squared:  0.8244
## F-statistic:  63.9 on 5 and 62 DF,  p-value: < 2.2e-16
```

```r
#Pruning of Variables to arrive at Final Model
slm_gam2 <- lm(formula = gmv ~ Promotion + cat_mid + sale_days + Sponsorship +
                 NPS, data = gam_train)
summary(slm_gam2)
```

```
##
## Call:
## lm(formula = gmv ~ Promotion + cat_mid + sale_days + Sponsorship +
##     NPS, data = gam_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1544179  -440047   -44635   343456  2379851
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21240855    2626057   8.088 2.78e-11 ***
## Promotion    5622738    2314864   2.429   0.0181 *
## cat_mid     -6284504    1306841  -4.809 1.01e-05 ***
## sale_days      96290      55152   1.746   0.0858 .
## Sponsorship   -39262       7690  -5.106 3.37e-06 ***
## NPS          -382603      49635  -7.708 1.27e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 760900 on 62 degrees of freedom
## Multiple R-squared:  0.8375, Adjusted R-squared:  0.8244
## F-statistic:  63.9 on 5 and 62 DF,  p-value: < 2.2e-16
```

```
vif(slm_gam2)
```

```
##   Promotion     cat_mid   sale_days Sponsorship         NPS
##   49.513168   49.968920    1.118252    5.654426    5.488590
```

```
#Removing Promotion
slm_gam3 <- lm(formula = gmv ~ cat_mid + sale_days + Sponsorship +
                 NPS, data = gam_train)
summary(slm_gam3)
```

```
##
## Call:
## lm(formula = gmv ~ cat_mid + sale_days + Sponsorship + NPS, data = gam_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2554139  -457993   -36163   458403  2702941
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21445470    2724865   7.870 6.02e-11 ***
## cat_mid     -3143149     194895 -16.127  < 2e-16 ***
## sale_days     108019      57037   1.894   0.0628 .
## Sponsorship   -33218       7554  -4.398 4.30e-05 ***
## NPS          -344691      48916  -7.047 1.66e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 790000 on 63 degrees of freedom
## Multiple R-squared:  0.822,  Adjusted R-squared:  0.8107
## F-statistic: 72.75 on 4 and 63 DF,  p-value: < 2.2e-16
```

```
vif(slm_gam3)
```

```
##     cat_mid   sale_days Sponsorship         NPS
##    1.031163    1.109680    5.062429    4.945853
```

```
#Removing sale_days, This is our final model
slm_gam4 <- lm(formula = gmv ~ cat_mid + Sponsorship +
                 NPS, data = gam_train)
summary(slm_gam4)
```

```
##
## Call:
## lm(formula = gmv ~ cat_mid + Sponsorship + NPS, data = gam_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2586545  -388669    -5954   484926  2881107
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 20233899    2701697   7.489 2.56e-10 ***
## cat_mid     -3138603     198779 -15.789  < 2e-16 ***
## Sponsorship   -28815       7331  -3.931 0.000211 ***
## NPS          -321987      48373  -6.656 7.43e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 805800 on 64 degrees of freedom
## Multiple R-squared:  0.8119, Adjusted R-squared:  0.8031
## F-statistic: 92.08 on 3 and 64 DF,  p-value: < 2.2e-16
```

```
vif(slm_gam4)
```

```
##     cat_mid Sponsorship         NPS
##    1.031006    4.582699    4.648777
```

```
#Test the model on Test Dataset
pred_gam_slm<-predict(slm_gam4,gam_test[,-1])

#Add New column for predicted_gmv
gam_test$predicted_gmv <- pred_gam_slm

#Lets look at the Corr & R2
cor(gam_test$gmv, gam_test$predicted_gmv)
```

```
## [1] 0.9172916
```

```
cor(gam_test$gmv, gam_test$predicted_gmv)^2
```

```
## [1] 0.8414238
```

```
#xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
```

Initial Linear Model

```
slm_hom1 <- lm(gmv~ .,data=hom_train)
```

Auto-Optimize Model

```
step_slm_hom <- stepAIC(slm_hom1, direction = "both",trace=FALSE)
summary(step_slm_hom)
```

```
##
## Call:
## lm(formula = gmv ~ product_mrp + list_price + Promotion + sla +
##     sale_days + ContentMarketing + Affiliates + SEM + NPS, data = hom_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3554718  -699727   -18358   524541  4781797
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.312e+07  7.731e+06  -2.990 0.004161 **
## product_mrp      -8.193e+02  8.116e+01 -10.094 3.99e-14 ***
## list_price        1.056e+03  1.211e+02   8.723 5.87e-12 ***
## Promotion         4.106e+07  5.119e+06   8.021 8.07e-11 ***
## sla               4.492e+05  1.417e+05   3.171 0.002486 **
## sale_days         1.993e+05  9.164e+04   2.175 0.033983 *
```

```
## ContentMarketing -2.139e+06  4.129e+05  -5.181 3.24e-06 ***
## Affiliates        3.749e+06  1.042e+06   3.598 0.000688 ***
## SEM               3.142e+05  7.599e+04   4.134 0.000123 ***
## NPS               1.825e+05  1.225e+05   1.489 0.142170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1257000 on 55 degrees of freedom
## Multiple R-squared:  0.8802, Adjusted R-squared:  0.8606
## F-statistic: 44.92 on 9 and 55 DF,  p-value: < 2.2e-16
```

```
#Pruning of Variables to arrive at Final Model
slm_hom2 <- lm(formula = gmv ~ product_mrp + list_price + Promotion + sla +
                sale_days + ContentMarketing + Affiliates + SEM + NPS, data = hom_train)
summary(slm_hom2)
```

```
##
## Call:
## lm(formula = gmv ~ product_mrp + list_price + Promotion + sla +
##     sale_days + ContentMarketing + Affiliates + SEM + NPS, data = hom_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3554718  -699727   -18358   524541  4781797
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.312e+07  7.731e+06  -2.990 0.004161 **
## product_mrp      -8.193e+02  8.116e+01 -10.094 3.99e-14 ***
## list_price        1.056e+03  1.211e+02   8.723 5.87e-12 ***
## Promotion         4.106e+07  5.119e+06   8.021 8.07e-11 ***
## sla               4.492e+05  1.417e+05   3.171 0.002486 **
## sale_days         1.993e+05  9.164e+04   2.175 0.033983 *
## ContentMarketing -2.139e+06  4.129e+05  -5.181 3.24e-06 ***
## Affiliates        3.749e+06  1.042e+06   3.598 0.000688 ***
## SEM               3.142e+05  7.599e+04   4.134 0.000123 ***
## NPS               1.825e+05  1.225e+05   1.489 0.142170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1257000 on 55 degrees of freedom
## Multiple R-squared:  0.8802, Adjusted R-squared:  0.8606
## F-statistic: 44.92 on 9 and 55 DF,  p-value: < 2.2e-16
```

```
vif(slm_hom2)
```

```
##      product_mrp       list_price        Promotion              sla
##       110.771699        74.470303        38.583116         1.637458
##        sale_days ContentMarketing       Affiliates              SEM
##         1.121092         7.022290        15.413117         2.728861
##              NPS
##         8.372622
```

```
cor(hom_train$product_mrp, hom_train$list_price)
```

```
## [1] 0.811916
```

```r
#Removing NPS & sale_day
slm_hom3 <- lm(formula = gmv ~ product_mrp + list_price + Promotion + sla +
                 ContentMarketing + Affiliates + SEM, data = hom_train)
summary(slm_hom3)
```

```
##
## Call:
## lm(formula = gmv ~ product_mrp + list_price + Promotion + sla +
##     ContentMarketing + Affiliates + SEM, data = hom_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3468891  -579098     2016   570011  5415889
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.234e+07  2.278e+06  -5.417 1.27e-06 ***
## product_mrp      -8.103e+02  8.304e+01  -9.758 9.17e-14 ***
## list_price        1.039e+03  1.240e+02   8.386 1.57e-11 ***
## Promotion         4.067e+07  5.235e+06   7.769 1.66e-10 ***
## sla               4.369e+05  1.470e+05   2.972 0.004324 **
## ContentMarketing -1.651e+06  3.642e+05  -4.534 3.02e-05 ***
## Affiliates        2.369e+06  6.352e+05   3.729 0.000444 ***
## SEM               2.240e+05  5.722e+04   3.915 0.000244 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1313000 on 57 degrees of freedom
## Multiple R-squared:  0.8647, Adjusted R-squared:  0.8481
## F-statistic: 52.05 on 7 and 57 DF,  p-value: < 2.2e-16
```

```r
vif(slm_hom3)
```

```
##      product_mrp       list_price        Promotion              sla
##       106.388220        71.579192        37.019528         1.617479
## ContentMarketing       Affiliates              SEM
##         5.011777         5.254915         1.419595
```

```r
#Removing sla
slm_hom4 <- lm(formula = gmv ~ product_mrp + list_price + Promotion +
                 ContentMarketing + Affiliates + SEM, data = hom_train)
summary(slm_hom4)
```

```
##
## Call:
## lm(formula = gmv ~ product_mrp + list_price + Promotion + ContentMarketing +
##     Affiliates + SEM, data = hom_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4928079  -462906    15371   471626  5669352
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -7.763e+06  1.789e+06  -4.341 5.76e-05 ***
```

```
## product_mrp      -7.588e+02  8.653e+01  -8.770 3.19e-12 ***
## list_price        9.685e+02  1.296e+02   7.474 4.68e-10 ***
## Promotion         3.677e+07  5.399e+06   6.810 6.11e-09 ***
## ContentMarketing -1.360e+06  3.737e+05  -3.639 0.000584 ***
## Affiliates        1.536e+06  6.074e+05   2.529 0.014183 *
## SEM               2.366e+05  6.080e+04   3.891 0.000260 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1398000 on 58 degrees of freedom
## Multiple R-squared:  0.8438, Adjusted R-squared:  0.8276
## F-statistic: 52.21 on 6 and 58 DF,  p-value: < 2.2e-16
```

```
vif(slm_hom4)
```

```
##      product_mrp       list_price        Promotion ContentMarketing
##       101.759099        68.924444        34.689557         4.649076
##       Affiliates              SEM
##         4.232430         1.411822
```

```
#Removing product_mrp
slm_hom5 <- lm(formula = gmv ~ list_price + Promotion +
                 ContentMarketing + Affiliates + SEM, data = hom_train)
summary(slm_hom5)
```

```
##
## Call:
## lm(formula = gmv ~ list_price + Promotion + ContentMarketing +
##     Affiliates + SEM, data = hom_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5152652 -1124960  -389520  1273799  8421017
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.828e+06  9.921e+05   6.882 4.28e-09 ***
## list_price       -1.592e+02  2.415e+01  -6.594 1.31e-08 ***
## Promotion        -9.865e+06  1.414e+06  -6.979 2.94e-09 ***
## ContentMarketing -1.247e+06  5.647e+05  -2.207  0.03120 *
## Affiliates        2.800e+06  8.922e+05   3.139  0.00265 **
## SEM               1.543e+05  9.083e+04   1.699  0.09460 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2114000 on 59 degrees of freedom
## Multiple R-squared:  0.6366, Adjusted R-squared:  0.6058
## F-statistic: 20.67 on 5 and 59 DF,  p-value: 7.201e-12
```

```
vif(slm_hom5)
```

```
##       list_price        Promotion ContentMarketing       Affiliates
##         1.046727         1.040018         4.643510         3.993956
##              SEM
##         1.378212
```

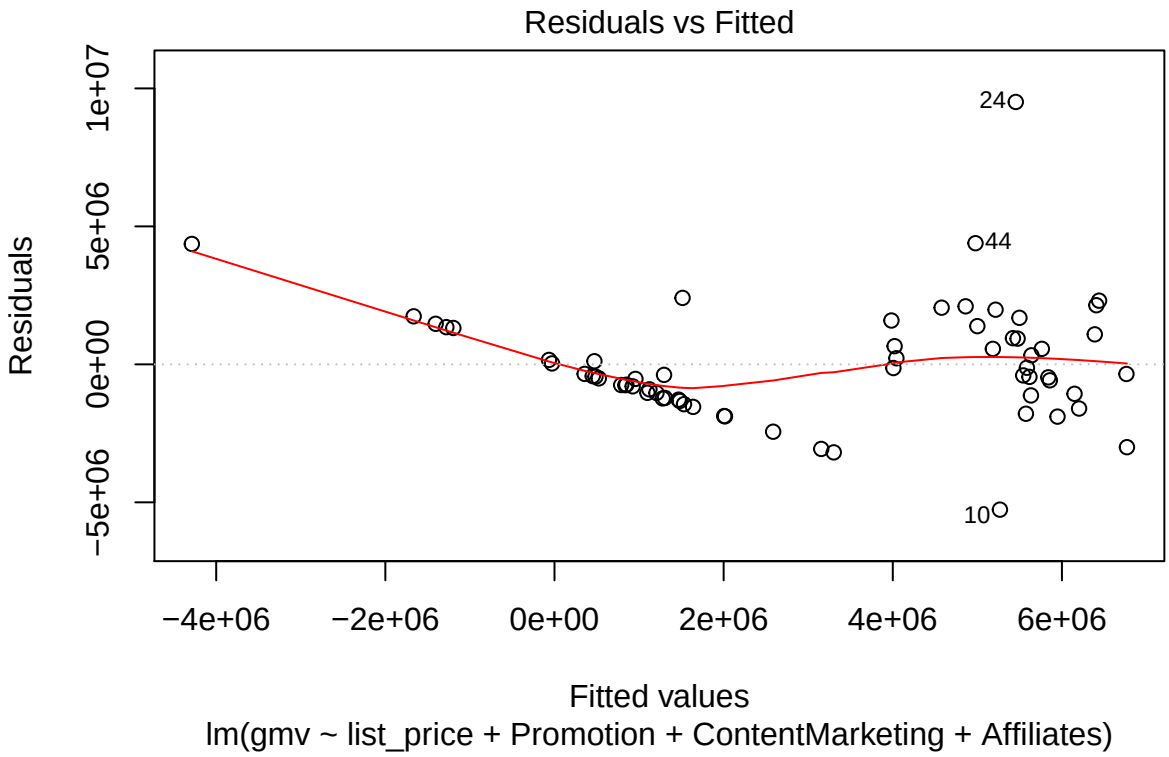
```r
#Removing SEM
slm_hom6 <- lm(formula = gmv ~ list_price + Promotion +
                 ContentMarketing + Affiliates, data = hom_train)
summary(slm_hom6)
```

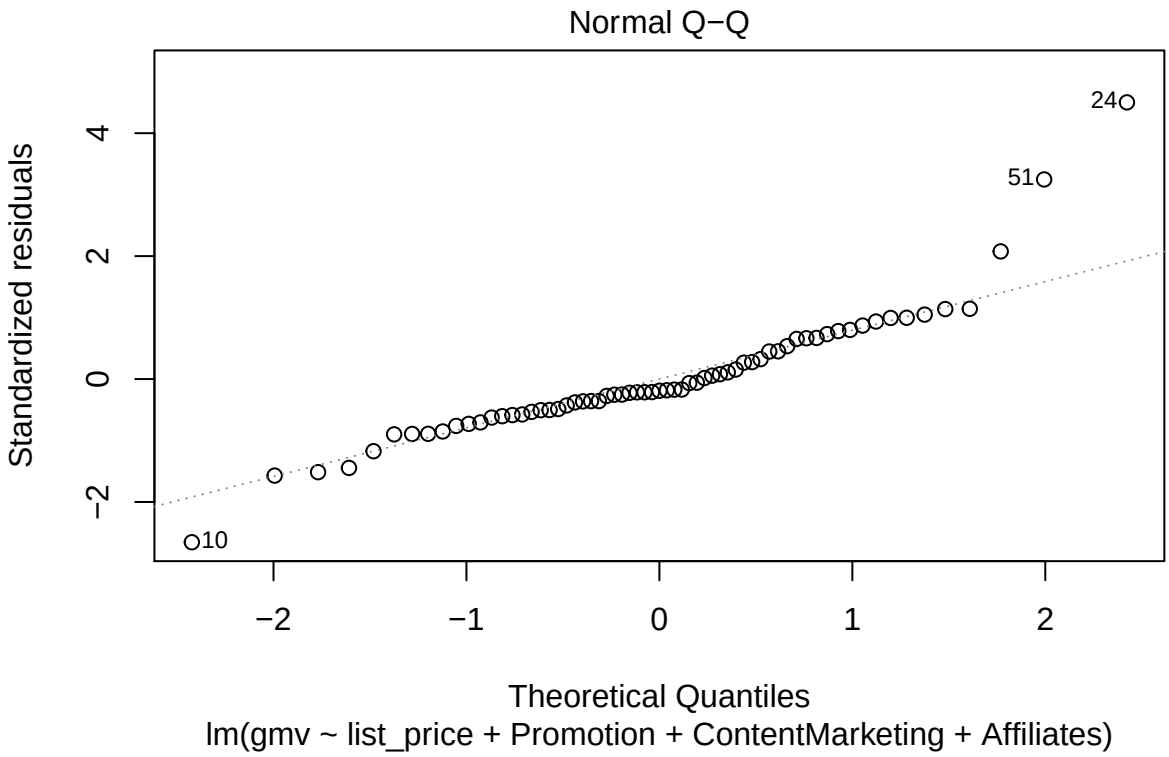
```
##
## Call:
## lm(formula = gmv ~ list_price + Promotion + ContentMarketing +
##     Affiliates, data = hom_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5264514 -1126480  -401283  1087356  9508494
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7212881.6   980971.2   7.353 6.25e-10 ***
## list_price           -161.2       24.5  -6.581 1.29e-08 ***
## Promotion        -9977176.0  1434042.1  -6.957 2.96e-09 ***
## ContentMarketing  -891469.0   532835.1  -1.673  0.09952 .
## Affiliates        2630310.3   900381.1   2.921  0.00491 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2147000 on 60 degrees of freedom
## Multiple R-squared:  0.6188, Adjusted R-squared:  0.5934
## F-statistic: 24.35 on 4 and 60 DF,  p-value: 5.313e-12
```

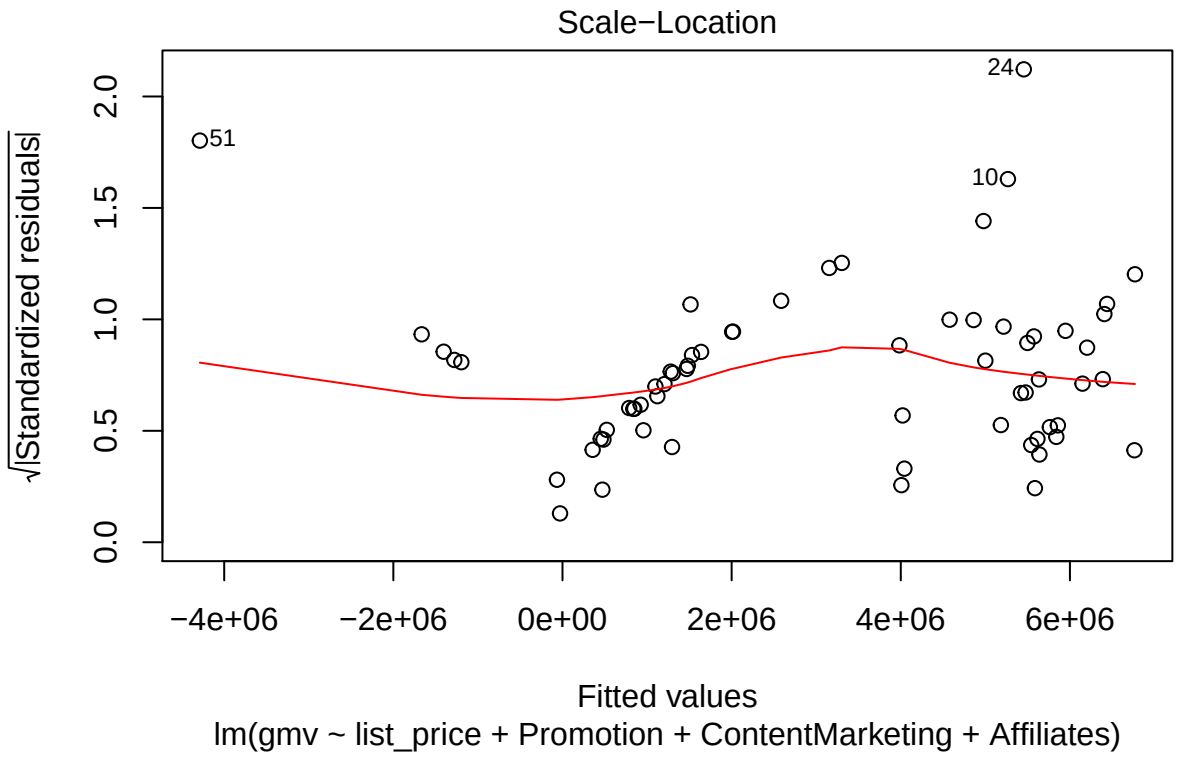
```r
vif(slm_hom6)
```

```
##       list_price        Promotion ContentMarketing       Affiliates
##         1.044310         1.037759         4.007606         3.943692
```

```r
plot(slm_hom6)
```

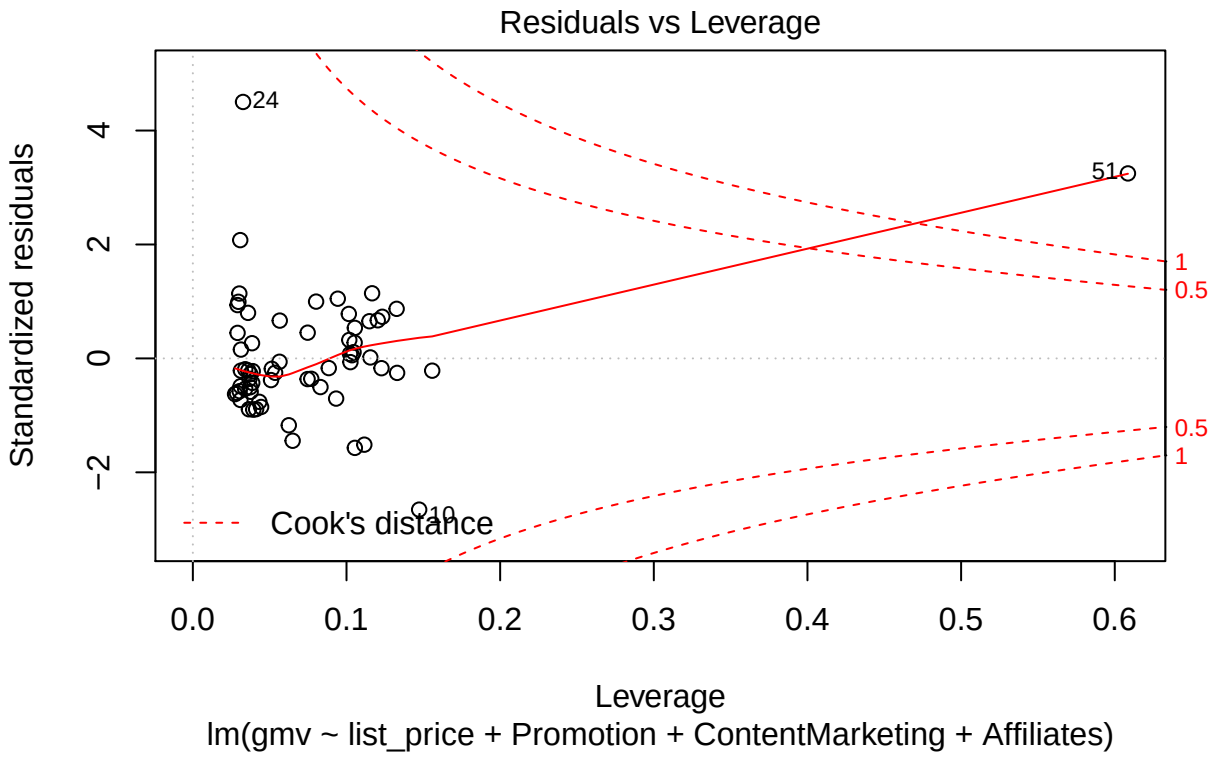

Residuals vs Fitted
Residuals
1e+07
5e+06
0e+00
−5e+06
24
44
10
−4e+06
−2e+06
0e+00
2e+06
4e+06
6e+06
Fitted values
lm(gmv ~ list_price + Promotion + ContentMarketing + Affiliates)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(gmv ~ list_price + Promotion + ContentMarketing + Affiliates)
24
51
10
−2
0
2
4
−2
−1
0
1
2




```
abline(slm_hom6)
```

```
## Warning in abline(slm_hom6): only using the first two of 5 regression
## coefficients
```




```
#Test the model on Test Dataset
pred_hom_slm<-predict(slm_hom6,hom_test[,-1])

#Add New column for predicted_gmv
hom_test$predicted_gmv <- pred_hom_slm

#Lets look at the Corr & R2
cor(hom_test$gmv, hom_test$predicted_gmv)
```

```
## [1] 0.8683034
```

```
cor(hom_test$gmv, hom_test$predicted_gmv)^2
```

```
## [1] 0.7539507
```