MarketMixModeling Models

Atchireddy chavva

Sat May 06 08:43:01 2017

Load Libraries:

Load required libraries. Will use `stepAIC` from `MASS` package for model pruning.

```
library(MASS)
library(stargazer)
```

Load Data:

Load blended data from sales and marketing datasets. Data is throughly cleaned, pre-processed for model building. Refer to `DataCleaning.R` script for data preparation steps. For this capstone project we limit our model building focus to `camera_accessory`, `Home_audio` and `Gaming_accessory` product sub-categories. For simplicity will start Linear model building with numerical features, later will consider categorical features.

```
camera_accessory_data_nrm <- read.csv('./intrim/cameraAccessory.csv')
home_audio_data_nrm       <- read.csv('./intrim/homeAudio.csv')
gaming_accessory_data_nrm <- read.csv('./intrim/gamingAccessory')
```

Lets preview the dataset structure. `gmv` gross merchandise values is our target variable, which we would like to maximize with optimal marketing spend across different marketing levers, `discount` is one of the KPI derived from sales data, bunch of other features from marketing spend and NPS datasets.

```
str(camera_accessory_data_nrm)
```

```
## 'data.frame':    52 obs. of  14 variables:
##  $ gmv             : num  18196 3341843 4884098 4514154 3588231 ...
##  $ discount        : num  1.562 -0.276 -0.791 -0.477 -0.397 ...
##  $ sla             : num  -4.772 0.9 0.682 0.302 0.284 ...
##  $ procurement_sla : num  0.101 0.1938 0.0836 0.4656 0.2653 ...
##  $ TV              : num  -1.42 -1.39 -1.39 -1.39 -1.4 ...
##  $ Digital         : num  -0.34291 0.00364 0.00364 0.00364 -0.05181 ...
##  $ Sponsorship     : num  -1.077 -0.948 -0.948 -0.948 -0.984 ...
##  $ ContentMarketing: num  -0.756 -0.756 -0.756 -0.756 -0.756 ...
##  $ OnlineMarketing : num  -1.94 -1.87 -1.87 -1.87 -1.89 ...
##  $ Affiliates      : num  -2 -1.92 -1.92 -1.92 -1.94 ...
##  $ SEM             : num  -0.634 -0.349 -0.349 -0.349 -0.397 ...
##  $ Radio           : num  -0.523 -0.523 -0.523 -0.523 -0.523 ...
##  $ Other           : num  -0.511 -0.511 -0.511 -0.511 -0.511 ...
##  $ NPS             : num  1.28 1.28 1.28 1.28 1.96 ...
```

Features distribution: Look at features statistical distributions

```
stargazer(camera_accessory_data_nrm,type='text')
```

```
##
## ========================================================================
## Statistic        N      Mean        St. Dev.       Min          Max
## ------------------------------------------------------------------------
## gmv             52 5,509,740.000 2,481,745.000 1,397.000 13,102,157.000
## discount        52     0.000         1.000      -3.126       4.616
## sla             52    -0.000         1.000      -4.772       3.169
```

```
## procurement_sla  52     0.000         1.000      -4.130       4.646
## TV               52     0.000         1.000      -1.468       2.197
## Digital          52     0.000         1.000      -0.643       3.271
## Sponsorship      52    -0.000         1.000      -1.204       2.198
## ContentMarketing 52     0.000         1.000      -0.756       3.047
## OnlineMarketing  52    -0.000         1.000      -2.019       1.025
## Affiliates       52     0.000         1.000      -2.081       0.936
## SEM              52     0.000         1.000      -0.681       3.221
## Radio            52     0.000         1.000      -0.523       3.004
## Other            52    -0.000         1.000      -0.511       2.876
## NPS              52     0.000         1.000      -1.280       2.640
## ---------------------------------------------------------------------------
```

Train and Validation Data:

With one year Sales and Marketing Data, we have 52 obeservations aggreagted at weekly level for each sub-category. Splitting data into training and validation sets would further reduce the training sample size, Moreover the task at hand is to figure out most influential marketing leavers for optimial marketing spend. The goal is to explain the influence of features rather predicting any quantities we can safely utilize the whole dataset for training.

Model Building - Linear Model:

Assumptions:

For simplicity, will consider, each sub-category sales affected from overall marketing spend, where in reality, only a portion of the marketing spend would have been alloted for promoting a certain product category.

Camera Accesory:

Initial Linear Model

```
model_cam1 <- lm(gmv~ .,data=camera_accessory_data_nrm)
```

Auto-Optimize Model

```
step_cam <- stepAIC(model_cam1, direction = "both",trace=FALSE,k=2)
```

Summary

```
stargazer(model_cam1,step_cam, align = TRUE, type = 'text',
          title='Linear Regression Results', single.row=TRUE)
```

```
##
## Linear Regression Results
## ======================================================================================
##                                            Dependent variable:
##                      -----------------------------------------------------------------
##                                                    gmv
##                                    (1)                                (2)
## --------------------------------------------------------------------------------------
## discount                13,381.270 (252,059.500)
## sla                    352,086.700 (319,485.100)       341,307.700 (244,386.300)
## procurement_sla        354,552.400 (260,324.500)       383,522.900 (237,266.500)
## TV                   -2,369,619.000 (4,157,446.000)  -2,660,457.000*** (736,201.100)
## Digital               7,654,649.000 (6,868,969.000)   8,070,851.000*** (2,799,984.000)
## Sponsorship         4,467,389.000*** (1,448,393.000)  4,218,906.000*** (804,623.800)
## ContentMarketing     -3,536,650.000 (4,422,996.000)  -4,920,915.000** (2,060,112.000)
## OnlineMarketing      -5,320,773.000 (5,744,324.000)
## Affiliates            8,122,930.000 (6,602,404.000)   3,725,745.000*** (871,406.200)
## SEM                 -6,451,448.000 (10,803,616.000)  -5,877,756.000** (2,676,528.000)
## Radio                 3,182,333.000 (8,454,713.000)   1,670,248.000*** (607,088.400)
## Other                -1,638,532.000 (9,560,871.000)
## NPS                    -209,026.500 (1,314,502.000)
## Constant              5,509,740.000*** (235,273.200)   5,509,740.000*** (227,063.500)
## --------------------------------------------------------------------------------------
## Observations                     52                                 52
## R2                              0.652                              0.642
## Adjusted R2                     0.533                              0.565
## Residual Std. Error     1,696,579.000 (df = 38)         1,637,378.000 (df = 42)
## F Statistic              5.471*** (df = 13; 38)           8.351*** (df = 9; 42)
## ======================================================================================
## Note:                                                      *p<0.1; **p<0.05; ***p<0.01
```

Understanding Model:

Linear model could explain 56% of revenue from marketing expenditure, but some of the significant features like `TV spending` has negative coefficient term. If we were to explain this, it should mean, `reducing the TV marketing spend would  increase camera Accessory sales`, which doesn't make sense. we can further optimize model by exploring multi-collinearity and pruning features. Hopefully we may end up with a model with better R-Square value and co-efficient terms which makes sense.

Gaming Accesory:

Initial Model

```
model_ga1 <- lm(gmv~ .,data=gaming_accessory_data_nrm)
```

Auto-Optimize Model

```
step_ga <- stepAIC(model_ga1, direction = "both",trace=FALSE)
```

```
stargazer(model_ga1,step_ga, align = TRUE, type = 'text',
          title='Linear Regression Results', single.row=TRUE)
```

```
##
## Linear Regression Results
## ==========================================================================================
##                                               Dependent variable:
##                       --------------------------------------------------------------------
##                                                       gmv
##                                     (1)                                  (2)
## ------------------------------------------------------------------------------------------
## discount               492,264.400*** (170,894.200)     447,874.400*** (159,252.100)
## sla                      -68,255.220 (120,702.100)
## procurement_sla          201,412.100 (132,045.200)        206,101.300 (130,092.400)
## TV                     -1,653,248.000 (1,660,227.000)   -2,593,057.000*** (446,675.800)
## Digital               7,341,960.000** (2,772,307.000) 5,801,867.000*** (1,438,317.000)
## Sponsorship            3,605,580.000*** (629,880.700)   3,721,704.000*** (522,787.300)
## ContentMarketing      -4,868,625.000** (1,838,756.000) -5,671,379.000*** (1,112,845.000)
## OnlineMarketing        -4,191,629.000 (2,585,858.000)   -4,126,524.000 (2,493,308.000)
## Affiliates            6,995,738.000** (2,884,398.000) 7,450,677.000*** (2,458,982.000)
## SEM                    -4,732,325.000 (4,237,833.000)   -2,232,922.000* (1,308,610.000)
## Radio                  4,004,619.000 (3,298,355.000)    1,896,536.000*** (319,929.600)
## Other                  -2,336,728.000 (3,716,188.000)
## NPS                      520,617.300 (525,942.800)       724,563.100** (330,032.800)
## Constant               3,229,366.000*** (102,965.000)   3,229,366.000*** (101,676.600)
## ------------------------------------------------------------------------------------------
## Observations                        53                                   53
## R2                                 0.794                                0.789
## Adjusted R2                        0.726                                0.733
## Residual Std. Error       749,596.400 (df = 39)              740,216.900 (df = 41)
## F Statistic              11.590*** (df = 13; 39)            13.955*** (df = 11; 41)
## ==========================================================================================
## Note:                                                         *p<0.1; **p<0.05; ***p<0.01
```

Home Accessory

Initial Model

```
model_ha1 <- lm(gmv~ .,data=home_audio_data_nrm)
```

Auto-Opitimize Model

```
step_ha <- stepAIC(model_ha1, direction = "both",trace=FALSE)
```

Summary

```
stargazer(model_ha1,step_ha, align = TRUE, type = 'text',
          title='Linear Regression Results', single.row=TRUE)
```

```
##
## Linear Regression Results
## ========================================================================================
##                                              Dependent variable:
##                     --------------------------------------------------------------------
##                                                      gmv
##                                    (1)                                (2)
## ----------------------------------------------------------------------------------------
## discount              2,904,574.000*** (433,630.000)     2,944,063.000*** (412,513.100)
## sla                   1,104,916.000*** (381,902.200)     1,157,871.000*** (343,981.000)
## procurement_sla       -980,968.000*** (275,741.400)      -987,332.200*** (271,781.700)
## TV                    5,141,356.000 (4,299,518.000)      6,109,515.000* (3,163,836.000)
## Digital              18,481,940.000** (7,612,086.000)  19,884,266.000*** (6,300,980.000)
## Sponsorship           2,746,332.000* (1,382,489.000)     2,394,255.000** (896,178.400)
## ContentMarketing      5,585,090.000 (4,840,766.000)      6,387,531.000 (4,165,723.000)
## OnlineMarketing       -5,610,540.000 (4,991,805.000)    -4,123,055.000* (2,314,829.000)
## Affiliates            1,964,885.000 (5,822,385.000)
## SEM                 -26,003,966.000** (12,087,329.000) -28,405,803.000*** (9,652,125.000)
## Radio                19,317,311.000* (9,534,118.000)    20,993,554.000** (8,039,993.000)
## Other               -19,951,450.000* (10,701,862.000)  -21,935,345.000** (8,834,894.000)
## NPS                   -1,105,701.000 (1,176,216.000)     -1,358,797.000 (895,183.600)
## Constant              5,345,857.000*** (228,263.800)     5,345,857.000*** (225,513.900)
## ----------------------------------------------------------------------------------------
## Observations                      50                                 50
## R2                              0.748                              0.747
## Adjusted R2                     0.657                              0.665
## Residual Std. Error      1,614,069.000 (df = 36)            1,594,624.000 (df = 37)
## F Statistic               8.205*** (df = 13; 36)             9.097*** (df = 12; 37)
## ========================================================================================
## Note:                                                      *p<0.1; **p<0.05; ***p<0.01
```

Observations: