MarketMixModeling DataPreparation

Atchireddy chavva

Thu May 11 23:26:58 2017

Goal:

ElecKart is an e-commerce firm specialising in electronic products. Over the last one year, they had spent a significant amount of money in marketing. Occasionally, they had also offered big-ticket promotions (similar to the Big Billion Day). They are about to create a marketing budget for the next year which includes spending on commercials, online campaigns, and pricing & promotion strategies. The CFO feels that the money spent over last 12 months on marketing was not sufficiently impactful and that they can either cut on the budget or reallocate it optimally across marketing levers to improve the revenue response

```
# ****************************************************************************
#                     LOAD LIBRARY ----
# ****************************************************************************
```

```
library(lubridate)
library(dplyr)
library(ggplot2)
library(MASS)
library(car)
library(Hmisc)   # describe


# ****************************************************************************
#                     PROCs ----
# ****************************************************************************
# Function to compute the week number w.r.t origin date.
# It takes data and orgin in Date format as arguments.
nweek <- function(x, format="%Y-%m-%d", origin){
  if(missing(origin)){
    as.integer(format(strptime(x, format=format), "%W"))
  }else{
    x <- as.Date(x, format=format)
    o <- as.Date(origin, format=format)
    w <- as.integer(format(strptime(x, format=format), "%w"))
    2 + as.integer(x - o - w) %/% 7
  }
}
```

```r
# ****************************************************************************
#                    LOAD DATA ---- Transaction Data ----
# ****************************************************************************
# Make sure you are in current directory as in R-file is in. Should I do a commit?yes..

ce_data <- read.csv('./input/ConsumerElectronics.csv',stringsAsFactors = FALSE)
```

```r
str(ce_data)
```

```
## 'data.frame':    1648824 obs. of  20 variables:
##  $ ï..fsn_id                     : chr  "ACCCX3S58G7B5F6P" "ACCCX3S58G7B5F6P" "ACCCX3S5AHMF55FV" "AC
##  $ order_date                    : chr  "2015-10-17 15:11:54" "2015-10-19 10:07:22" "2015-10-20 15:4
##  $ Year                          : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ Month                         : int  10 10 10 10 10 10 10 10 10 10 ...
##  $ order_id                      : num  3.42e+15 1.42e+15 2.42e+15 4.42e+15 4.42e+15 ...
##  $ order_item_id                 : num  3.42e+15 1.42e+15 2.42e+15 4.42e+15 4.42e+15 ...
##  $ gmv                           : num  6400 6900 1990 1690 1618 ...
##  $ units                         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ deliverybdays                 : chr  "\\N" "\\N" "\\N" "\\N" ...
##  $ deliverycdays                 : chr  "\\N" "\\N" "\\N" "\\N" ...
##  $ s1_fact.order_payment_type    : chr  "COD" "COD" "COD" "Prepaid" ...
##  $ sla                           : int  5 7 10 4 6 5 6 5 9 7 ...
##  $ cust_id                       : num  -1.01e+18 -8.99e+18 -1.04e+18 -7.60e+18 2.89e+18 ...
##  $ pincode                       : num  -7.79e+18 7.34e+18 -7.48e+18 -5.84e+18 5.35e+17 ...
##  $ product_analytic_super_category: chr  "CE" "CE" "CE" "CE" ...
##  $ product_analytic_category     : chr  "CameraAccessory" "CameraAccessory" "CameraAccessory" "Came
##  $ product_analytic_sub_category : chr  "CameraAccessory" "CameraAccessory" "CameraAccessory" "Came
##  $ product_analytic_vertical     : chr  "CameraTripod" "CameraTripod" "CameraTripod" "CameraTripod"
##  $ product_mrp                   : int  7190 7190 2099 2099 2099 4044 4044 4044 4044 4044 ...
##  $ product_procurement_sla       : int  0 0 3 3 3 5 5 5 5 5 ...
```

```r
# ****************************************************************************
#                    DATA CLEANING ----
# ****************************************************************************


# . . . .   Outlier Treatment ----
# Remove orders before July'15 and after June'16
ce_data$order_date <- format(as.POSIXct(ce_data$order_date,format='%Y-%m-%d'),
                             format='%Y-%m-%d')
ce_data$order_date <- as.Date(ce_data$order_date, format = "%Y-%m-%d")

ce_data <- subset(ce_data, order_date > "2015-6-30" & order_date < "2016-7-1")


# . . . .   Missing Values ----

# Since 80% of the variable has NAs Omit 'deliverybday' & 'deliverycdays'
# Removed Pincode, as there seems to be some data quality issues with this variable
ce_data <- ce_data[,-c(9,10)]

ce_data <- na.omit(ce_data)   # 4904 missing values, can be ignored
```

```r
# . . . .   Correct Data Types ----

# 'order_id', 'order_item_id', 'cust_id', 'pincode' are qualitative data
#  having numeric values, let's convert them to character type

ce_data <- cbind(ce_data[,-c(5,6,11,12)],
                 sapply(ce_data[,c(5,6,11,12)],as.character) )
# operate on interested columns



# ****************************************************************************
#                    FEATURE ENGINEERING ----
# ****************************************************************************

# create week,  week numbers start from min 'order date'
# . . . . Week Numbers ----
dates <- as.Date(
  gsub(" .*","",ce_data$order_date)
)
ce_data$week <- nweek(dates,origin = as.Date("2015-07-01"))


# . . . . Days, weeks, Month ----
# will compute Month, week, and no.of days per week (month, week)
#
dys <- seq(as.Date("2015-07-01"),as.Date("2016-06-30"),'days')
weekdays <- data.frame('days'=dys, Month = month(dys),
                       week = nweek(dys,origin = as.Date("2015-07-01")),
                       nweek = rep(1,length(dys)))
weekdays <- data.frame(weekdays %>%
                         group_by(Month,week) %>%
                              summarise(nweeks = sum(nweek)))
weekdays$fracDays <- weekdays$nweeks/7


# . . . . Strip Spaces ----
ce_data$product_analytic_vertical <- gsub(" +","",ce_data$product_analytic_vertical)


# . . . . Generate Discount ----
ce_data$discount <- ((ce_data$product_mrp - ce_data$gmv)/ce_data$product_mrp) * 100

# . . . .  Payment Type ----
ce_data$COD     = as.integer(ce_data$s1_fact.order_payment_type=='COD')
ce_data$Prepaid = as.integer(ce_data$s1_fact.order_payment_type!='COD')
```

```r
# ********************************************************************************
#                    LOAD DATA ---- Media & Inv Data ----
# ********************************************************************************
# . . . .   ProductList ----
productList_data      <-
  read.csv("./input/ProductList.csv", stringsAsFactors = FALSE,
           na.strings=c('\\N'))

# . . . .   Media Investment ----
mediaInvestment_data  <-
  read.csv("./input/MediaInvestment.csv", stringsAsFactors = FALSE)

# . . . .  Special Sale Event ----

specialSale_data      <-
  read.csv("./input/SpecialSale.csv", stringsAsFactors = FALSE)

# . . . .   Monthly NPS ----
monthlyNPS_data       <-
  read.csv("./input/MonthlyNPSscore.csv", stringsAsFactors = FALSE )




# ********************************************************************************
#                    DATA PREPARATION ----
# ********************************************************************************

# . . . .   ProductList ----
productList_data <- na.omit(productList_data)

# . . . . . .  Correct Data types ----
productList_data$Frequency <- as.integer(productList_data$Frequency)
```

```r
str(productList_data)
```

```
## 'data.frame':    73 obs. of  3 variables:
##  $ Product  : chr  "AmplifierReceiver" "AudioMP3Player" "Binoculars" "BoomBox" ...
##  $ Frequency: int  4056 112892 14599 2879 987 2269 17523 41307 15660 401 ...
##  $ Percent  : num  0.2 6.8 0.9 0.2 0.1 0.1 1.1 2.5 0.9 0 ...
##  - attr(*, "na.action")=Class 'omit'  Named int 1
##   .. ..- attr(*, "names")= chr "1"
```

```r
# . . . .   Media Investment ----
str(mediaInvestment_data)
```

```
## 'data.frame':    12 obs. of  12 variables:
##  $ Year             : int  2015 2015 2015 2015 2015 2015 2016 2016 2016 2016 ...
##  $ Month            : int  7 8 9 10 11 12 1 2 3 4 ...
##  $ Total.Investment : num  17.1 5.1 96.3 170.2 51.2 ...
##  $ TV               : num  0.2 0 3.9 6.1 4.2 5.4 4.4 2.6 9.3 5.2 ...
##  $ Digital          : num  2.5 1.3 1.4 12.6 1.3 3.1 0.5 1.9 2.1 0.9 ...
##  $ Sponsorship      : num  7.4 1.1 62.8 84.7 14.2 56.7 4.2 11.7 41.6 24.3 ...
```

```
##  $ Content.Marketing: num  0 0 0.6 3.4 0.2 1.1 0.9 0.6 0.4 0 ...
##  $ Online.marketing : num  1.3 0.1 16.4 24.4 19.6 22.5 22.9 19.9 18.4 16.5 ...
##  $ Affiliates       : num  0.5 0.1 5 7 6.6 6.8 7.4 6.5 6.2 5.7 ...
##  $ SEM              : num  5 2.5 6.2 31.9 5.2 11.2 4.2 4.9 5.2 4.2 ...
##  $ Radio            : num  NA NA NA NA NA NA 2.7 NA 0.9 NA ...
##  $ Other            : num  NA NA NA NA NA NA 27.1 NA 15.9 NA ...
```

```
# . . . . . .  Missing Values ----
mediaInvestment_data[is.na(mediaInvestment_data)] <- 0   # zero investment


# . . . . . .  Convert to weekly data ----
# convert montly spend to weekly
mediaInvestment_data <- cbind(Month=mediaInvestment_data[,c(2)],
                              mediaInvestment_data[,-c(1,2)]/4.30)
# Add weekly information
mediaInvestment_weekly <- merge(weekdays,mediaInvestment_data, by='Month',
                                all.x = TRUE)

# Convert media Investment at weekly granularity
# pro-rate weekly investment as per the ratio of its days span over adjacent months
mediaInvestment_weekly <-
  data.frame(
            mediaInvestment_weekly %>%
                group_by(week) %>%
                    summarise(TotalInvestment = sum(Total.Investment*fracDays),
                              TV = sum(TV*fracDays),
                              Digital=sum(Digital*fracDays),
                              Sponsorship = sum(Sponsorship*fracDays),
                              ContentMarketing = sum(Content.Marketing*fracDays),
                              OnlineMarketing = sum(Online.marketing*fracDays),
                              Affiliates = sum(Affiliates*fracDays),
                              SEM = sum(SEM*fracDays),
                              Radio = sum(Radio*fracDays),
                              Other = sum(Other*fracDays))
                            )

# . . . .   SPecialSale ----
str(specialSale_data)
```

```
## 'data.frame':    44 obs. of  2 variables:
##  $ Day         : chr  "7/18/2015" "7/19/2015" "8/15/2015" "8/16/2015" ...
##  $ SaleOccasion: chr  "Eid_RathaYatraSale" "Eid_RathaYatraSale" "IndependenceSale" "IndependenceSale
```

```
specialSale_data$Date      <- as.Date(specialSale_data$Day, format = "%m/%d/%Y")
specialSale_data$week      <- nweek(specialSale_data$Date,origin = as.Date("2015-07-01"))
specialSale_data           <- data.frame(table(specialSale_data$week))
colnames(specialSale_data) <- c('week','n_saledays')


# . . . .   Monthly NPS ----
monthlyNPS_weekly   <- merge(weekdays, monthlyNPS_data, by='Month', all.x = TRUE)
monthlyNPS_weekly   <- as.data.frame(monthlyNPS_weekly %>% group_by(., week) %>%
                                           summarise(., NPS = mean(NPS)))
```

```
# ************************************************************************
#                   WEEKLY DATA AGGREGATION ----
# ************************************************************************
ce_data_weekly <-  ce_data %>%
  group_by(product_analytic_sub_category,
           week) %>%
  summarise(gmv=sum(gmv),
            product_mrp=mean(product_mrp),
            units=sum(units),
            discount=mean(discount),
            sla=mean(sla),
            procurement_sla=mean(product_procurement_sla),
            COD=sum(COD),
            Prepaid = sum(Prepaid))

ce_data_weekly <- as.data.frame(ce_data_weekly)   # type cast to data.frame



# ************************************************************************
#                   MERGING DATA ----
# ************************************************************************

# . . . .   Merge MediaInvestment & NPS ----
media_nps <- merge(mediaInvestment_weekly, monthlyNPS_weekly, by = 'week',
                   all.x = TRUE)


# . . . .   Merge Sales & SaleDays
data <- merge(ce_data_weekly, specialSale_data, by = 'week', all.x = TRUE)
data[is.na(data$n_saledays),'n_saledays'] = 0
# data$Sales.Name[is.na(data$Sales.Name)] <- "No sale"


#. . . .   Merge Data & Media_NPS
data <- merge(data, media_nps, by = 'week', all.x = TRUE)


# Discount on Products
data$list_mrp     <- as.integer(data$gmv/data$units)
data$discount     <- (1-(data$list_mrp/data$product_mrp))*100


# . . . . Save Intrim Data ----
write.csv(data, file = "./intrim/eleckart.csv",row.names=FALSE)
```

```
str(data)
```

```
## 'data.frame':    667 obs. of  23 variables:
##  $ week                          : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ product_analytic_sub_category: chr  "AmplifierReceiver" "Camera" "AudioAccessory" "CameraAccessory
##  $ gmv                           : num  4766 105340 12849 18196 26313 ...
##  $ product_mrp                   : num  2375 15995 7211 1423 2326 ...
```

```
##  $ units                     : int  4 9 3 28 17 6 24 33 67 5 ...
##  $ discount                  : num  49.8 26.8 40.6 54.4 33.5 ...
##  $ sla                       : num  4.75 3.11 6.33 2.5 3.06 ...
##  $ procurement_sla           : num  3.25 2.44 1.67 2.61 2.19 ...
##  $ COD                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Prepaid                   : int  4 9 3 28 16 5 24 32 54 5 ...
##  $ n_saledays                : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TotalInvestment           : num  2.27 2.27 2.27 2.27 2.27 ...
##  $ TV                        : num  0.0266 0.0266 0.0266 0.0266 0.0266 ...
##  $ Digital                   : num  0.332 0.332 0.332 0.332 0.332 ...
##  $ Sponsorship               : num  0.983 0.983 0.983 0.983 0.983 ...
##  $ ContentMarketing          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ OnlineMarketing           : num  0.173 0.173 0.173 0.173 0.173 ...
##  $ Affiliates                : num  0.0664 0.0664 0.0664 0.0664 0.0664 ...
##  $ SEM                       : num  0.664 0.664 0.664 0.664 0.664 ...
##  $ Radio                     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Other                     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ NPS                       : num  54.6 54.6 54.6 54.6 54.6 54.6 54.6 54.6 54.6 54.6 ...
##  $ list_mrp                  : int  1191 11704 4283 649 1547 493 2951 1197 1649 7323 ...
```