Tayko Software Cataloger

Raghav Arora

2023-05-07

##Assignment of Tayko

Tayko.com <- read.csv("Tayko.csv",header=TRUE)  
head(Tayko.com)

## sequence\_number US source\_a source\_c source\_b source\_d source\_e source\_m  
## 1 1 1 0 0 1 0 0 0  
## 2 2 1 0 0 0 0 1 0  
## 3 3 1 0 0 0 0 0 0  
## 4 4 1 0 1 0 0 0 0  
## 5 5 1 0 1 0 0 0 0  
## 6 6 1 0 0 0 0 0 0  
## source\_o source\_h source\_r source\_s source\_t source\_u source\_p source\_x  
## 1 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 1 0 0 0  
## 4 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0  
## 6 0 0 1 0 0 0 0 0  
## source\_w Freq last\_update\_days\_ago X1st\_update\_days\_ago Web.order Gender.male  
## 1 0 2 3662 3662 1 0  
## 2 0 0 2900 2900 1 1  
## 3 0 2 3883 3914 0 0  
## 4 0 1 829 829 0 1  
## 5 0 1 869 869 0 0  
## 6 0 1 1995 2002 0 0  
## Address\_is\_res Purchase Spending  
## 1 1 1 128  
## 2 0 0 0  
## 3 0 1 127  
## 4 0 0 0  
## 5 0 0 0  
## 6 1 0 0

dim(Tayko.com)

## [1] 2000 25

1. Each catalog costs approximately $2 to mail (including printing, postage, and mailing costs). Estimate the gross profit that the firm could expect from the remaining 180,000 names if it selects them randomly from the pool.

#gross\_profit <- cost\_per\_mail \* response\_rate \* average\_spend \* number\_of\_left\_entries  
cost\_per\_mail <- 2  
response\_rate <- 0.053  
number\_of\_left\_entries <- 180000  
average\_spend <- mean(Tayko.com[Tayko.com$Purchase==1,]$Spending)  
gross\_profit <- (average\_spend \* response\_rate - cost\_per\_mail) \* number\_of\_left\_entries  
gross\_profit

## [1] 1598075

1. Develop a model for classifying a customer as a purchaser or non-purchaser.
2. Partition the data randomly into a training set (800 records), validation set (700 records), and test set (500 records).

#Droped last\_update\_days\_ago & 1st\_update\_days\_ago but kept sequence\_number required as purchaseID  
Tayko.com <- Tayko.com[,-c(19,20)]  
set.seed(1)  
train\_rows <- sample(rownames(Tayko.com),dim(Tayko.com)[1]\*0.4)  
valid\_rows <- sample(setdiff(rownames(Tayko.com),train\_rows),dim(Tayko.com)[1]\*0.35)  
test\_rows <- setdiff(rownames(Tayko.com),union(train\_rows,valid\_rows))  
train.data <- Tayko.com[train\_rows,] # 800  
valid.data <- Tayko.com[valid\_rows,] # 700  
test.data <- Tayko.com[test\_rows,] # 500

1. Run step-wise logistic regression using backward elimination to select the best subset of variables, then use this model to classify the data into purchasers and non-purchasers. Use only the training set for running the model. (Logistic regression is used because it yields an estimated “probability of purchase,” which is required later in the analysis.)

#Drop spending from variables  
lr\_model <- glm(Purchase ~ .-Spending, data = train.data, family = binomial);

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

lr\_model\_b <- step(lr\_model, direction = "backward");

## Start: AIC=662.64  
## Purchase ~ (sequence\_number + US + source\_a + source\_c + source\_b +   
## source\_d + source\_e + source\_m + source\_o + source\_h + source\_r +   
## source\_s + source\_t + source\_u + source\_p + source\_x + source\_w +   
## Freq + Web.order + Gender.male + Address\_is\_res + Spending) -   
## Spending

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - source\_c 1 618.64 660.64  
## - source\_b 1 618.70 660.70  
## - source\_o 1 618.72 660.72  
## - source\_t 1 619.17 661.17  
## - US 1 619.26 661.26  
## - source\_d 1 619.32 661.32  
## - source\_m 1 619.43 661.43  
## - Gender.male 1 619.47 661.47  
## - sequence\_number 1 619.57 661.57  
## - source\_r 1 619.68 661.68  
## - source\_e 1 620.56 662.56  
## <none> 618.64 662.64  
## - source\_x 1 621.51 663.51  
## - source\_s 1 621.57 663.57  
## - source\_p 1 622.07 664.07  
## - source\_w 1 622.25 664.25  
## - source\_a 1 624.75 666.75  
## - Address\_is\_res 1 627.18 669.18  
## - source\_h 1 627.30 669.30  
## - source\_u 1 628.97 670.97  
## - Web.order 1 649.76 691.76  
## - Freq 1 948.80 990.80

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=660.64  
## Purchase ~ sequence\_number + US + source\_a + source\_b + source\_d +   
## source\_e + source\_m + source\_o + source\_h + source\_r + source\_s +   
## source\_t + source\_u + source\_p + source\_x + source\_w + Freq +   
## Web.order + Gender.male + Address\_is\_res

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - source\_o 1 618.75 658.75  
## - source\_b 1 618.77 658.77  
## - US 1 619.26 659.26  
## - Gender.male 1 619.47 659.47  
## - sequence\_number 1 619.57 659.57  
## - source\_t 1 619.59 659.59  
## - source\_d 1 619.83 659.83  
## - source\_m 1 619.83 659.83  
## <none> 618.64 660.64  
## - source\_r 1 621.10 661.10  
## - source\_x 1 622.73 662.73  
## - source\_p 1 623.18 663.18  
## - source\_e 1 623.85 663.85  
## - source\_s 1 624.68 664.68  
## - Address\_is\_res 1 627.18 667.18  
## - source\_w 1 628.43 668.43  
## - source\_a 1 634.31 674.31  
## - source\_h 1 638.09 678.09  
## - source\_u 1 644.86 684.86  
## - Web.order 1 649.79 689.79  
## - Freq 1 950.65 990.65

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=658.75  
## Purchase ~ sequence\_number + US + source\_a + source\_b + source\_d +   
## source\_e + source\_m + source\_h + source\_r + source\_s + source\_t +   
## source\_u + source\_p + source\_x + source\_w + Freq + Web.order +   
## Gender.male + Address\_is\_res

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - source\_b 1 618.93 656.93  
## - US 1 619.36 657.36  
## - Gender.male 1 619.56 657.56  
## - source\_t 1 619.62 657.62  
## - sequence\_number 1 619.62 657.62  
## - source\_m 1 619.86 657.86  
## - source\_d 1 620.07 658.07  
## <none> 618.75 658.75  
## - source\_r 1 621.10 659.10  
## - source\_x 1 622.74 660.74  
## - source\_p 1 623.18 661.18  
## - source\_e 1 623.95 661.95  
## - source\_s 1 624.73 662.73  
## - Address\_is\_res 1 627.31 665.31  
## - source\_w 1 628.78 666.78  
## - source\_a 1 634.88 672.88  
## - source\_h 1 639.54 677.54  
## - source\_u 1 646.13 684.13  
## - Web.order 1 649.85 687.85  
## - Freq 1 951.50 989.50

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=656.93  
## Purchase ~ sequence\_number + US + source\_a + source\_d + source\_e +   
## source\_m + source\_h + source\_r + source\_s + source\_t + source\_u +   
## source\_p + source\_x + source\_w + Freq + Web.order + Gender.male +   
## Address\_is\_res

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - US 1 619.58 655.58  
## - Gender.male 1 619.68 655.68  
## - sequence\_number 1 619.83 655.83  
## - source\_d 1 620.09 656.09  
## - source\_t 1 620.13 656.13  
## - source\_m 1 620.30 656.30  
## <none> 618.93 656.93  
## - source\_r 1 622.23 658.23  
## - source\_x 1 623.42 659.42  
## - source\_p 1 623.88 659.88  
## - source\_s 1 626.32 662.32  
## - source\_e 1 626.35 662.35  
## - Address\_is\_res 1 627.43 663.43  
## - source\_w 1 632.96 668.96  
## - source\_a 1 639.96 675.96  
## - source\_h 1 641.05 677.05  
## - Web.order 1 650.14 686.14  
## - source\_u 1 654.30 690.30  
## - Freq 1 951.51 987.51

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=655.58  
## Purchase ~ sequence\_number + source\_a + source\_d + source\_e +   
## source\_m + source\_h + source\_r + source\_s + source\_t + source\_u +   
## source\_p + source\_x + source\_w + Freq + Web.order + Gender.male +   
## Address\_is\_res

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - Gender.male 1 620.38 654.38  
## - sequence\_number 1 620.59 654.59  
## - source\_t 1 620.65 654.65  
## - source\_d 1 620.71 654.71  
## - source\_m 1 621.00 655.00  
## <none> 619.58 655.58  
## - source\_r 1 622.90 656.90  
## - source\_x 1 624.17 658.17  
## - source\_p 1 624.39 658.39  
## - source\_e 1 626.49 660.49  
## - source\_s 1 627.23 661.23  
## - Address\_is\_res 1 627.83 661.83  
## - source\_w 1 633.02 667.02  
## - source\_a 1 640.67 674.67  
## - source\_h 1 642.08 676.08  
## - Web.order 1 650.73 684.73  
## - source\_u 1 654.35 688.35  
## - Freq 1 954.60 988.60

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=654.38  
## Purchase ~ sequence\_number + source\_a + source\_d + source\_e +   
## source\_m + source\_h + source\_r + source\_s + source\_t + source\_u +   
## source\_p + source\_x + source\_w + Freq + Web.order + Address\_is\_res

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - sequence\_number 1 621.37 653.37  
## - source\_t 1 621.38 653.38  
## - source\_d 1 621.44 653.44  
## - source\_m 1 621.70 653.70  
## <none> 620.38 654.38  
## - source\_r 1 623.72 655.72  
## - source\_x 1 624.93 656.93  
## - source\_p 1 625.45 657.45  
## - source\_e 1 627.24 659.24  
## - source\_s 1 627.80 659.80  
## - Address\_is\_res 1 628.41 660.41  
## - source\_w 1 633.82 665.82  
## - source\_a 1 641.54 673.54  
## - source\_h 1 642.67 674.67  
## - Web.order 1 651.45 683.45  
## - source\_u 1 654.90 686.90  
## - Freq 1 955.41 987.41

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=653.37  
## Purchase ~ source\_a + source\_d + source\_e + source\_m + source\_h +   
## source\_r + source\_s + source\_t + source\_u + source\_p + source\_x +   
## source\_w + Freq + Web.order + Address\_is\_res

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - source\_d 1 622.36 652.36  
## - source\_t 1 622.48 652.48  
## - source\_m 1 622.78 652.78  
## <none> 621.37 653.37  
## - source\_r 1 624.63 654.63  
## - source\_x 1 625.86 655.86  
## - source\_p 1 626.45 656.45  
## - source\_e 1 628.27 658.27  
## - source\_s 1 628.64 658.64  
## - Address\_is\_res 1 629.61 659.61  
## - source\_w 1 634.64 664.64  
## - source\_a 1 642.74 672.74  
## - source\_h 1 643.48 673.48  
## - Web.order 1 651.54 681.54  
## - source\_u 1 656.43 686.43  
## - Freq 1 957.00 987.00

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=652.36  
## Purchase ~ source\_a + source\_e + source\_m + source\_h + source\_r +   
## source\_s + source\_t + source\_u + source\_p + source\_x + source\_w +   
## Freq + Web.order + Address\_is\_res

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - source\_t 1 623.89 651.89  
## - source\_m 1 624.12 652.12  
## <none> 622.36 652.36  
## - source\_r 1 626.66 654.66  
## - source\_x 1 627.34 655.34  
## - source\_p 1 627.99 655.99  
## - Address\_is\_res 1 630.34 658.34  
## - source\_s 1 630.91 658.91  
## - source\_e 1 631.29 659.29  
## - source\_w 1 638.90 666.90  
## - source\_h 1 643.48 671.48  
## - source\_a 1 647.65 675.65  
## - Web.order 1 652.78 680.78  
## - source\_u 1 663.04 691.04  
## - Freq 1 957.10 985.10

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=651.89  
## Purchase ~ source\_a + source\_e + source\_m + source\_h + source\_r +   
## source\_s + source\_u + source\_p + source\_x + source\_w + Freq +   
## Web.order + Address\_is\_res

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - source\_m 1 625.28 651.28  
## <none> 623.89 651.89  
## - source\_r 1 627.27 653.27  
## - source\_x 1 628.30 654.30  
## - source\_p 1 628.97 654.97  
## - source\_s 1 631.38 657.38  
## - source\_e 1 631.49 657.49  
## - Address\_is\_res 1 631.62 657.62  
## - source\_w 1 638.91 664.91  
## - source\_a 1 647.66 673.66  
## - source\_h 1 648.67 674.67  
## - Web.order 1 655.57 681.57  
## - source\_u 1 663.24 689.24  
## - Freq 1 961.57 987.57

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=651.28  
## Purchase ~ source\_a + source\_e + source\_h + source\_r + source\_s +   
## source\_u + source\_p + source\_x + source\_w + Freq + Web.order +   
## Address\_is\_res

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## <none> 625.28 651.28  
## - source\_r 1 628.08 652.08  
## - source\_x 1 629.35 653.35  
## - source\_p 1 630.04 654.04  
## - source\_e 1 631.95 655.95  
## - source\_s 1 632.10 656.10  
## - Address\_is\_res 1 632.73 656.73  
## - source\_w 1 639.08 663.08  
## - source\_a 1 647.76 671.76  
## - source\_h 1 652.60 676.60  
## - Web.order 1 656.45 680.45  
## - source\_u 1 663.24 687.24  
## - Freq 1 964.69 988.69

library(caret)

## Warning: package 'caret' was built under R version 4.2.3

## Loading required package: ggplot2

## Loading required package: lattice

summary(lr\_model\_b)

##   
## Call:  
## glm(formula = Purchase ~ source\_a + source\_e + source\_h + source\_r +   
## source\_s + source\_u + source\_p + source\_x + source\_w + Freq +   
## Web.order + Address\_is\_res, family = binomial, data = train.data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.4657 -0.5538 -0.1717 0.6083 2.2843   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.0812 0.3519 -11.599 < 2e-16 \*\*\*  
## source\_a 1.6844 0.3657 4.606 4.11e-06 \*\*\*  
## source\_e 0.8689 0.3376 2.574 0.010051 \*   
## source\_h -2.8871 0.6357 -4.542 5.58e-06 \*\*\*  
## source\_r 0.7035 0.4187 1.680 0.092908 .   
## source\_s 1.2393 0.4668 2.655 0.007931 \*\*   
## source\_u 2.1005 0.3588 5.854 4.81e-09 \*\*\*  
## source\_p 2.2378 1.1745 1.905 0.056743 .   
## source\_x 1.6178 0.7757 2.086 0.037023 \*   
## source\_w 1.1635 0.3175 3.664 0.000248 \*\*\*  
## Freq 2.2837 0.1925 11.862 < 2e-16 \*\*\*  
## Web.order 1.1071 0.2028 5.459 4.79e-08 \*\*\*  
## Address\_is\_res -0.7349 0.2718 -2.704 0.006843 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1108.96 on 799 degrees of freedom  
## Residual deviance: 625.28 on 787 degrees of freedom  
## AIC: 651.28  
##   
## Number of Fisher Scoring iterations: 6

prediction <- predict(lr\_model\_b, train.data, type = "response")  
classification <- ifelse(prediction > 0.5, 1, 0)  
confusionMatrix(factor(classification),factor(train.data$Purchase))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 334 73  
## 1 70 323  
##   
## Accuracy : 0.8212   
## 95% CI : (0.7929, 0.8472)  
## No Information Rate : 0.505   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.6424   
##   
## Mcnemar's Test P-Value : 0.8672   
##   
## Sensitivity : 0.8267   
## Specificity : 0.8157   
## Pos Pred Value : 0.8206   
## Neg Pred Value : 0.8219   
## Prevalence : 0.5050   
## Detection Rate : 0.4175   
## Detection Prevalence : 0.5088   
## Balanced Accuracy : 0.8212   
##   
## 'Positive' Class : 0   
##

1. Develop a model for predicting spending among the purchasers.
2. Create a vector of ID’s of only purchasers’ records (Purchase = 1).

# Only include purchase  
purchaser <- Tayko.com[Tayko.com$Purchase == 1, ]  
purchaser\_id <- purchaser$sequence\_number  
purchaser\_id

## [1] 1 3 9 10 14 15 19 21 22 24 25 26 29 30  
## [15] 32 37 42 43 44 45 46 49 51 53 55 57 58 59  
## [29] 61 63 64 65 66 68 69 71 77 78 79 81 82 84  
## [43] 86 87 89 90 94 96 100 105 106 108 109 111 113 114  
## [57] 115 118 119 120 121 122 123 125 128 129 130 131 132 133  
## [71] 136 138 140 143 146 147 148 149 150 152 156 159 160 164  
## [85] 166 167 168 169 170 172 178 179 180 182 183 185 186 187  
## [99] 188 189 190 191 192 193 194 195 198 201 202 204 207 208  
## [113] 209 210 211 212 213 214 217 219 220 221 223 224 227 228  
## [127] 229 232 233 234 235 236 237 238 239 240 241 246 247 248  
## [141] 249 250 253 255 257 259 260 262 263 265 266 267 268 269  
## [155] 270 271 272 276 277 279 280 282 283 284 286 288 289 290  
## [169] 292 293 295 297 299 300 301 302 303 304 305 308 309 310  
## [183] 311 313 315 317 318 319 320 321 323 324 327 331 332 335  
## [197] 341 344 345 346 347 349 350 351 352 354 355 362 365 368  
## [211] 369 370 371 373 374 376 377 378 379 380 381 384 385 386  
## [225] 387 388 389 390 391 393 394 400 401 403 404 405 407 409  
## [239] 410 412 418 420 423 424 425 427 428 431 432 433 437 438  
## [253] 439 440 442 445 446 447 448 449 450 453 454 455 456 457  
## [267] 458 459 461 462 467 471 476 477 481 484 486 487 488 490  
## [281] 492 493 494 496 498 499 505 508 511 516 518 519 520 522  
## [295] 526 527 528 533 534 537 538 539 541 542 543 545 547 551  
## [309] 553 555 556 559 561 562 566 567 568 569 571 574 575 576  
## [323] 577 578 579 580 582 583 585 586 587 588 589 590 592 596  
## [337] 597 599 601 602 603 607 612 615 618 621 623 624 625 627  
## [351] 628 634 635 636 641 643 644 645 646 649 650 651 653 656  
## [365] 664 665 666 667 668 674 676 678 681 682 687 690 691 693  
## [379] 695 696 697 698 702 703 704 705 707 708 712 716 717 721  
## [393] 722 724 726 729 730 733 734 735 736 737 739 740 743 744  
## [407] 745 746 750 753 754 755 761 762 765 768 769 775 776 777  
## [421] 778 779 780 783 786 787 788 792 793 799 802 803 804 805  
## [435] 807 812 813 816 820 821 823 826 828 829 830 835 837 839  
## [449] 840 843 846 850 851 852 853 854 855 856 862 863 867 868  
## [463] 869 871 873 875 877 880 882 887 888 893 894 895 896 898  
## [477] 899 902 905 906 907 909 912 915 917 918 919 923 924 926  
## [491] 927 929 930 933 934 937 946 948 949 950 952 954 955 958  
## [505] 959 960 962 964 965 967 968 969 973 975 977 979 980 981  
## [519] 983 984 987 989 992 993 994 996 998 999 1003 1006 1008 1009  
## [533] 1010 1014 1018 1020 1021 1024 1025 1026 1027 1028 1029 1032 1033 1035  
## [547] 1036 1037 1040 1042 1044 1045 1046 1047 1048 1049 1053 1055 1056 1057  
## [561] 1060 1062 1063 1064 1066 1068 1069 1072 1073 1076 1080 1084 1086 1087  
## [575] 1088 1089 1092 1098 1099 1100 1102 1104 1106 1107 1109 1110 1115 1116  
## [589] 1118 1126 1133 1136 1139 1142 1143 1144 1146 1147 1149 1153 1157 1159  
## [603] 1160 1162 1165 1166 1169 1171 1173 1175 1179 1180 1184 1187 1188 1189  
## [617] 1190 1193 1195 1201 1202 1206 1209 1215 1216 1218 1220 1224 1226 1227  
## [631] 1228 1231 1234 1236 1237 1238 1239 1241 1242 1243 1245 1247 1250 1252  
## [645] 1253 1256 1257 1259 1260 1261 1263 1264 1269 1270 1273 1274 1275 1276  
## [659] 1277 1279 1280 1286 1287 1289 1292 1294 1296 1298 1299 1300 1303 1305  
## [673] 1311 1313 1316 1319 1321 1322 1323 1324 1325 1328 1330 1331 1332 1334  
## [687] 1335 1336 1337 1339 1341 1343 1345 1350 1353 1354 1359 1361 1362 1363  
## [701] 1366 1368 1370 1371 1373 1375 1377 1380 1384 1386 1388 1389 1390 1391  
## [715] 1398 1399 1400 1402 1404 1420 1424 1426 1427 1430 1431 1433 1437 1440  
## [729] 1444 1447 1448 1450 1451 1453 1455 1456 1458 1459 1464 1465 1469 1470  
## [743] 1472 1474 1475 1476 1477 1478 1479 1485 1486 1488 1493 1494 1496 1499  
## [757] 1500 1501 1502 1506 1507 1516 1517 1519 1520 1521 1522 1523 1525 1526  
## [771] 1527 1528 1529 1531 1532 1533 1534 1537 1538 1539 1541 1542 1543 1544  
## [785] 1545 1546 1548 1549 1550 1551 1553 1554 1556 1559 1560 1561 1562 1563  
## [799] 1564 1566 1568 1572 1574 1576 1578 1580 1583 1584 1585 1587 1588 1589  
## [813] 1591 1594 1595 1597 1598 1599 1600 1602 1604 1605 1608 1610 1615 1616  
## [827] 1618 1619 1621 1624 1625 1628 1631 1632 1634 1640 1644 1648 1649 1654  
## [841] 1655 1656 1661 1665 1666 1668 1671 1672 1675 1677 1678 1679 1686 1688  
## [855] 1689 1690 1692 1693 1696 1697 1700 1701 1702 1704 1706 1707 1709 1711  
## [869] 1713 1714 1717 1720 1721 1724 1726 1727 1730 1734 1735 1740 1742 1744  
## [883] 1748 1751 1752 1753 1755 1758 1759 1760 1761 1763 1765 1766 1768 1775  
## [897] 1776 1777 1781 1782 1785 1786 1792 1794 1800 1802 1805 1808 1809 1810  
## [911] 1813 1814 1815 1818 1821 1824 1827 1829 1830 1832 1833 1836 1838 1839  
## [925] 1841 1845 1849 1852 1853 1855 1856 1861 1862 1864 1865 1867 1869 1872  
## [939] 1875 1876 1881 1883 1885 1888 1893 1894 1898 1900 1901 1902 1906 1907  
## [953] 1908 1910 1911 1915 1920 1922 1923 1926 1928 1929 1930 1932 1935 1940  
## [967] 1941 1946 1949 1950 1951 1954 1955 1957 1958 1959 1961 1962 1964 1966  
## [981] 1969 1970 1972 1973 1974 1975 1976 1977 1981 1986 1987 1988 1989 1990  
## [995] 1991 1992 1993 1995 1996 1997

1. Partition this dataset into the training and validation records. (Use the same training/validation labels from the earlier partitioning; one way is to use function intersect() to find IDs of purchasers in the original partitions).

train\_rows <- intersect(train.data$sequence\_number, Tayko.com[Tayko.com$Purchase == 1, ]$sequence\_number)  
valid\_rows <- intersect(valid.data$sequence\_number, Tayko.com[Tayko.com$Purchase == 1, ]$sequence\_number)  
train.data <- Tayko.com[Tayko.com$sequence\_number %in% train\_rows, ]  
valid.data <- Tayko.com[Tayko.com$sequence\_number %in% valid\_rows, ]

1. Develop models for predicting spending, using:
2. Multiple linear regression (use stepwise regression)

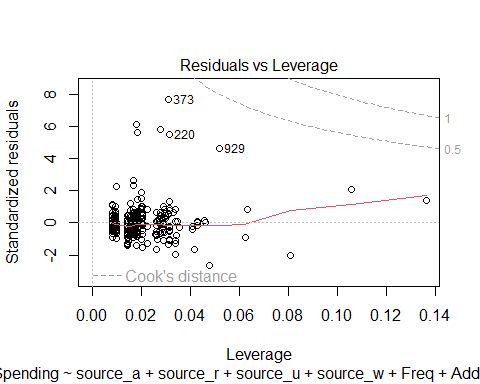
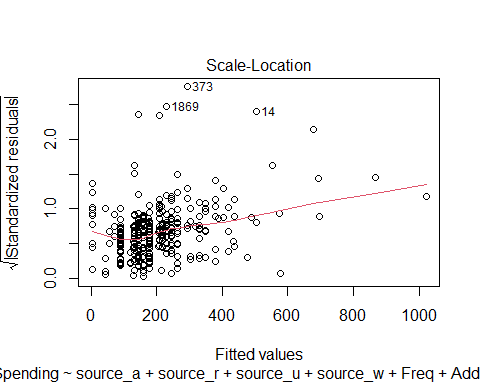
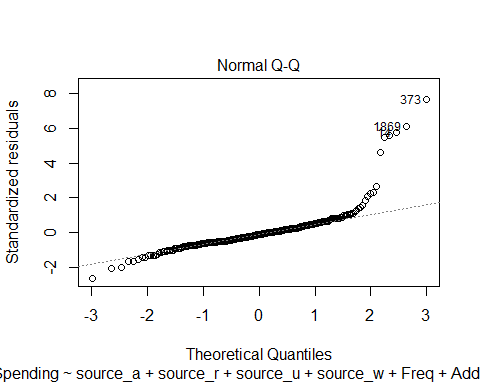
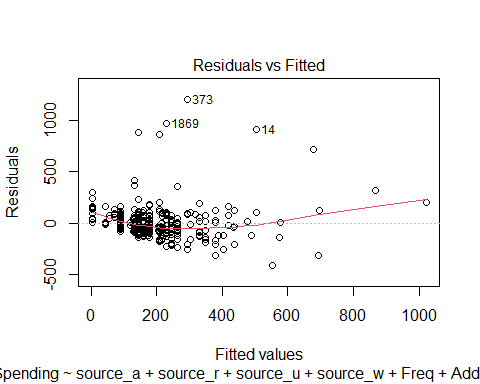
lm\_model <- lm(Spending ~ ., data = train.data)  
lm\_model <- step(lm\_model)

## Start: AIC=4102  
## Spending ~ sequence\_number + US + source\_a + source\_c + source\_b +   
## source\_d + source\_e + source\_m + source\_o + source\_h + source\_r +   
## source\_s + source\_t + source\_u + source\_p + source\_x + source\_w +   
## Freq + Web.order + Gender.male + Address\_is\_res + Purchase  
##   
##   
## Step: AIC=4102  
## Spending ~ sequence\_number + US + source\_a + source\_c + source\_b +   
## source\_d + source\_e + source\_m + source\_o + source\_h + source\_r +   
## source\_s + source\_t + source\_u + source\_p + source\_x + source\_w +   
## Freq + Web.order + Gender.male + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_t 1 10 11171639 4100.0  
## - Gender.male 1 60 11171689 4100.0  
## - sequence\_number 1 94 11171723 4100.0  
## - source\_o 1 200 11171829 4100.0  
## - source\_x 1 697 11172326 4100.0  
## - source\_p 1 1036 11172665 4100.0  
## - source\_s 1 1056 11172685 4100.0  
## - source\_c 1 1590 11173219 4100.1  
## - source\_u 1 3958 11175587 4100.1  
## - source\_a 1 4985 11176614 4100.2  
## - source\_b 1 5703 11177333 4100.2  
## - source\_r 1 5855 11177484 4100.2  
## - source\_e 1 8478 11180108 4100.3  
## - source\_d 1 9069 11180698 4100.3  
## - source\_m 1 12556 11184185 4100.4  
## - source\_h 1 14090 11185719 4100.5  
## - US 1 30257 11201886 4101.1  
## - Web.order 1 31036 11202665 4101.1  
## - source\_w 1 50128 11221757 4101.8  
## <none> 11171629 4102.0  
## - Address\_is\_res 1 415175 11586804 4114.4  
## - Freq 1 9211419 20383048 4338.1  
##   
## Step: AIC=4100  
## Spending ~ sequence\_number + US + source\_a + source\_c + source\_b +   
## source\_d + source\_e + source\_m + source\_o + source\_h + source\_r +   
## source\_s + source\_u + source\_p + source\_x + source\_w + Freq +   
## Web.order + Gender.male + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - Gender.male 1 59 11171698 4098.0  
## - sequence\_number 1 96 11171736 4098.0  
## - source\_o 1 207 11171846 4098.0  
## - source\_x 1 1130 11172769 4098.0  
## - source\_p 1 1648 11173287 4098.1  
## - source\_s 1 1669 11173308 4098.1  
## - source\_c 1 3379 11175018 4098.1  
## - source\_b 1 9922 11181561 4098.4  
## - source\_u 1 10515 11182154 4098.4  
## - source\_r 1 12668 11184307 4098.4  
## - source\_d 1 12887 11184526 4098.5  
## - source\_a 1 13180 11184819 4098.5  
## - source\_m 1 18481 11190120 4098.7  
## - source\_h 1 19820 11191459 4098.7  
## - source\_e 1 20912 11192552 4098.7  
## - US 1 30278 11201917 4099.1  
## - Web.order 1 31174 11202813 4099.1  
## <none> 11171639 4100.0  
## - source\_w 1 122002 11293641 4102.3  
## - Address\_is\_res 1 415732 11587371 4112.5  
## - Freq 1 9222275 20393914 4336.3  
##   
## Step: AIC=4098  
## Spending ~ sequence\_number + US + source\_a + source\_c + source\_b +   
## source\_d + source\_e + source\_m + source\_o + source\_h + source\_r +   
## source\_s + source\_u + source\_p + source\_x + source\_w + Freq +   
## Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - sequence\_number 1 103 11171801 4096.0  
## - source\_o 1 218 11171916 4096.0  
## - source\_x 1 1114 11172812 4096.0  
## - source\_s 1 1705 11173403 4096.1  
## - source\_p 1 1741 11173439 4096.1  
## - source\_c 1 3381 11175079 4096.1  
## - source\_b 1 9864 11181562 4096.4  
## - source\_u 1 10507 11182205 4096.4  
## - source\_r 1 12611 11184309 4096.4  
## - source\_d 1 12831 11184529 4096.5  
## - source\_a 1 13187 11184885 4096.5  
## - source\_m 1 18567 11190265 4096.7  
## - source\_h 1 20217 11191914 4096.7  
## - source\_e 1 20960 11192658 4096.7  
## - US 1 30337 11202035 4097.1  
## - Web.order 1 31410 11203107 4097.1  
## <none> 11171698 4098.0  
## - source\_w 1 122489 11294187 4100.3  
## - Address\_is\_res 1 415749 11587447 4110.5  
## - Freq 1 9249888 20421586 4334.9  
##   
## Step: AIC=4096.01  
## Spending ~ US + source\_a + source\_c + source\_b + source\_d + source\_e +   
## source\_m + source\_o + source\_h + source\_r + source\_s + source\_u +   
## source\_p + source\_x + source\_w + Freq + Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_o 1 256 11172057 4094.0  
## - source\_x 1 1092 11172893 4094.0  
## - source\_s 1 1724 11173524 4094.1  
## - source\_p 1 1725 11173526 4094.1  
## - source\_c 1 3392 11175193 4094.1  
## - source\_b 1 9974 11181775 4094.4  
## - source\_u 1 10548 11182349 4094.4  
## - source\_r 1 12570 11184370 4094.5  
## - source\_d 1 12803 11184604 4094.5  
## - source\_a 1 13139 11184940 4094.5  
## - source\_m 1 18574 11190375 4094.7  
## - source\_h 1 20170 11191970 4094.7  
## - source\_e 1 20933 11192734 4094.7  
## - US 1 30345 11202146 4095.1  
## - Web.order 1 32085 11203886 4095.1  
## <none> 11171801 4096.0  
## - source\_w 1 122401 11294202 4098.3  
## - Address\_is\_res 1 415800 11587601 4108.5  
## - Freq 1 9258654 20430455 4333.0  
##   
## Step: AIC=4094.01  
## Spending ~ US + source\_a + source\_c + source\_b + source\_d + source\_e +   
## source\_m + source\_h + source\_r + source\_s + source\_u + source\_p +   
## source\_x + source\_w + Freq + Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_x 1 1353 11173409 4092.1  
## - source\_s 1 1490 11173547 4092.1  
## - source\_p 1 2070 11174126 4092.1  
## - source\_c 1 4337 11176393 4092.2  
## - source\_b 1 9881 11181938 4092.4  
## - source\_d 1 12613 11184669 4092.5  
## - source\_u 1 13552 11185609 4092.5  
## - source\_r 1 15111 11187168 4092.6  
## - source\_a 1 16722 11188778 4092.6  
## - source\_m 1 18536 11190593 4092.7  
## - source\_h 1 21644 11193701 4092.8  
## - source\_e 1 25501 11197558 4092.9  
## - US 1 30860 11202917 4093.1  
## - Web.order 1 31857 11203913 4093.1  
## <none> 11172057 4094.0  
## - source\_w 1 142936 11314993 4097.0  
## - Address\_is\_res 1 415921 11587977 4106.5  
## - Freq 1 9263209 20435266 4331.1  
##   
## Step: AIC=4092.06  
## Spending ~ US + source\_a + source\_c + source\_b + source\_d + source\_e +   
## source\_m + source\_h + source\_r + source\_s + source\_u + source\_p +   
## source\_w + Freq + Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_p 1 1523 11174932 4090.1  
## - source\_s 1 2624 11176034 4090.2  
## - source\_c 1 3294 11176703 4090.2  
## - source\_u 1 12283 11185692 4090.5  
## - source\_b 1 13185 11186594 4090.5  
## - source\_r 1 13772 11187182 4090.6  
## - source\_d 1 15158 11188568 4090.6  
## - source\_a 1 15626 11189035 4090.6  
## - source\_h 1 20366 11193776 4090.8  
## - source\_m 1 22004 11195413 4090.8  
## - source\_e 1 24704 11198113 4090.9  
## - US 1 30213 11203622 4091.1  
## - Web.order 1 31135 11204545 4091.2  
## <none> 11173409 4092.1  
## - source\_w 1 154010 11327419 4095.5  
## - Address\_is\_res 1 414852 11588261 4104.5  
## - Freq 1 9264599 20438009 4329.2  
##   
## Step: AIC=4090.12  
## Spending ~ US + source\_a + source\_c + source\_b + source\_d + source\_e +   
## source\_m + source\_h + source\_r + source\_s + source\_u + source\_w +   
## Freq + Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_c 1 2416 11177348 4088.2  
## - source\_s 1 3888 11178820 4088.3  
## - source\_u 1 10760 11185693 4088.5  
## - source\_r 1 12297 11187229 4088.6  
## - source\_a 1 14119 11189051 4088.6  
## - source\_b 1 16407 11191339 4088.7  
## - source\_d 1 17360 11192292 4088.7  
## - source\_h 1 19150 11194082 4088.8  
## - source\_e 1 23357 11198289 4088.9  
## - source\_m 1 25018 11199950 4089.0  
## - US 1 31012 11205944 4089.2  
## - Web.order 1 31548 11206480 4089.2  
## <none> 11174932 4090.1  
## - source\_w 1 162361 11337293 4093.8  
## - Address\_is\_res 1 415027 11589959 4102.6  
## - Freq 1 9274006 20448938 4327.4  
##   
## Step: AIC=4088.2  
## Spending ~ US + source\_a + source\_b + source\_d + source\_e + source\_m +   
## source\_h + source\_r + source\_s + source\_u + source\_w + Freq +   
## Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_s 1 7012 11184360 4086.5  
## - source\_u 1 8347 11185695 4086.5  
## - source\_r 1 9929 11187277 4086.6  
## - source\_a 1 11834 11189182 4086.6  
## - source\_h 1 17223 11194571 4086.8  
## - source\_e 1 21380 11198727 4087.0  
## - source\_d 1 22080 11199428 4087.0  
## - source\_b 1 23813 11201161 4087.0  
## - US 1 29569 11206917 4087.2  
## - Web.order 1 31328 11208676 4087.3  
## - source\_m 1 31335 11208682 4087.3  
## <none> 11177348 4088.2  
## - source\_w 1 181396 11358744 4092.6  
## - Address\_is\_res 1 418888 11596236 4100.8  
## - Freq 1 9289161 20466509 4325.7  
##   
## Step: AIC=4086.45  
## Spending ~ US + source\_a + source\_b + source\_d + source\_e + source\_m +   
## source\_h + source\_r + source\_u + source\_w + Freq + Web.order +   
## Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_r 1 15863 11200223 4085.0  
## - source\_u 1 16721 11201081 4085.0  
## - source\_d 1 18830 11203190 4085.1  
## - source\_b 1 19302 11203662 4085.1  
## - source\_h 1 20408 11204768 4085.2  
## - source\_a 1 22151 11206511 4085.2  
## - source\_m 1 27344 11211704 4085.4  
## - Web.order 1 30999 11215359 4085.5  
## - US 1 31615 11215975 4085.6  
## - source\_e 1 33113 11217473 4085.6  
## <none> 11184360 4086.5  
## - source\_w 1 230095 11414455 4092.5  
## - Address\_is\_res 1 416605 11600965 4098.9  
## - Freq 1 9393780 20578140 4325.9  
##   
## Step: AIC=4085.01  
## Spending ~ US + source\_a + source\_b + source\_d + source\_e + source\_m +   
## source\_h + source\_u + source\_w + Freq + Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_u 1 7977 11208199 4083.3  
## - source\_a 1 12054 11212277 4083.4  
## - source\_h 1 15875 11216098 4083.6  
## - source\_e 1 22002 11222224 4083.8  
## - source\_d 1 24949 11225172 4083.9  
## - source\_b 1 29442 11229665 4084.1  
## - Web.order 1 31949 11232172 4084.1  
## - US 1 32614 11232837 4084.2  
## - source\_m 1 35552 11235775 4084.3  
## <none> 11200223 4085.0  
## - source\_w 1 215824 11416046 4090.6  
## - Address\_is\_res 1 405847 11606070 4097.1  
## - Freq 1 9402972 20603194 4324.4  
##   
## Step: AIC=4083.29  
## Spending ~ US + source\_a + source\_b + source\_d + source\_e + source\_m +   
## source\_h + source\_w + Freq + Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_a 1 6249 11214448 4081.5  
## - source\_h 1 13068 11221268 4081.8  
## - source\_e 1 15190 11223389 4081.8  
## - source\_d 1 30544 11238743 4082.4  
## - Web.order 1 33113 11241312 4082.5  
## - US 1 35454 11243653 4082.5  
## - source\_b 1 40048 11248247 4082.7  
## - source\_m 1 43047 11251246 4082.8  
## <none> 11208199 4083.3  
## - source\_w 1 219783 11427982 4089.0  
## - Address\_is\_res 1 406042 11614241 4095.4  
## - Freq 1 9395274 20603474 4322.4  
##   
## Step: AIC=4081.51  
## Spending ~ US + source\_b + source\_d + source\_e + source\_m + source\_h +   
## source\_w + Freq + Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_e 1 11370 11225819 4079.9  
## - source\_h 1 11537 11225985 4079.9  
## - Web.order 1 31298 11245746 4080.6  
## - US 1 33726 11248175 4080.7  
## - source\_d 1 34298 11248747 4080.7  
## - source\_b 1 46824 11261272 4081.2  
## - source\_m 1 47777 11262225 4081.2  
## <none> 11214448 4081.5  
## - source\_w 1 215394 11429842 4087.0  
## - Address\_is\_res 1 403849 11618297 4093.5  
## - Freq 1 9408658 20623106 4320.8  
##   
## Step: AIC=4079.92  
## Spending ~ US + source\_b + source\_d + source\_m + source\_h + source\_w +   
## Freq + Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_h 1 10014 11235832 4078.3  
## - Web.order 1 33481 11259300 4079.1  
## - source\_d 1 36590 11262408 4079.2  
## - US 1 40629 11266448 4079.3  
## - source\_m 1 51064 11276883 4079.7  
## - source\_b 1 51837 11277656 4079.7  
## <none> 11225819 4079.9  
## - source\_w 1 204124 11429943 4085.1  
## - Address\_is\_res 1 396153 11621971 4091.6  
## - Freq 1 9397426 20623245 4318.8  
##   
## Step: AIC=4078.27  
## Spending ~ US + source\_b + source\_d + source\_m + source\_w + Freq +   
## Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - Web.order 1 34956 11270788 4077.5  
## - source\_d 1 37067 11272900 4077.6  
## - US 1 42025 11277857 4077.7  
## - source\_m 1 52007 11287839 4078.1  
## - source\_b 1 53039 11288871 4078.1  
## <none> 11235832 4078.3  
## - source\_w 1 199704 11435536 4083.2  
## - Address\_is\_res 1 386239 11622071 4089.7  
## - Freq 1 9411071 20646904 4317.2  
##   
## Step: AIC=4077.5  
## Spending ~ US + source\_b + source\_d + source\_m + source\_w + Freq +   
## Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - US 1 39664 11310452 4076.9  
## - source\_d 1 40423 11311211 4076.9  
## - source\_b 1 50221 11321009 4077.3  
## - source\_m 1 50772 11321561 4077.3  
## <none> 11270788 4077.5  
## - source\_w 1 208113 11478901 4082.7  
## - Address\_is\_res 1 389749 11660537 4089.0  
## - Freq 1 9508684 20779472 4317.8  
##   
## Step: AIC=4076.89  
## Spending ~ source\_b + source\_d + source\_m + source\_w + Freq +   
## Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_d 1 45200 11355653 4076.5  
## - source\_b 1 55146 11365598 4076.8  
## - source\_m 1 56178 11366631 4076.9  
## <none> 11310452 4076.9  
## - source\_w 1 247575 11558027 4083.5  
## - Address\_is\_res 1 418869 11729321 4089.3  
## - Freq 1 9501295 20811747 4316.4  
##   
## Step: AIC=4076.47  
## Spending ~ source\_b + source\_m + source\_w + Freq + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_b 1 51478 11407131 4076.3  
## - source\_m 1 53647 11409300 4076.3  
## <none> 11355653 4076.5  
## - source\_w 1 258954 11614607 4083.4  
## - Address\_is\_res 1 408224 11763877 4088.5  
## - Freq 1 9457416 20813069 4314.4  
##   
## Step: AIC=4076.26  
## Spending ~ source\_m + source\_w + Freq + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_m 1 49866 11456997 4076.0  
## <none> 11407131 4076.3  
## - source\_w 1 279801 11686932 4083.9  
## - Address\_is\_res 1 397269 11804400 4087.8  
## - Freq 1 9412055 20819186 4312.5  
##   
## Step: AIC=4075.99  
## Spending ~ source\_w + Freq + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## <none> 11456997 4076.0  
## - source\_w 1 293517 11750514 4084.0  
## - Address\_is\_res 1 396478 11853475 4087.5  
## - Freq 1 9378299 20835296 4310.8

lm\_model <- lm(Spending ~ ., data = valid.data)  
lm\_model <- step(lm\_model)

## Start: AIC=3715.17  
## Spending ~ sequence\_number + US + source\_a + source\_c + source\_b +   
## source\_d + source\_e + source\_m + source\_o + source\_h + source\_r +   
## source\_s + source\_t + source\_u + source\_p + source\_x + source\_w +   
## Freq + Web.order + Gender.male + Address\_is\_res + Purchase  
##   
##   
## Step: AIC=3715.17  
## Spending ~ sequence\_number + US + source\_a + source\_c + source\_b +   
## source\_d + source\_e + source\_m + source\_o + source\_h + source\_r +   
## source\_s + source\_t + source\_u + source\_p + source\_x + source\_w +   
## Freq + Web.order + Gender.male + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_x 1 63 8956031 3713.2  
## - source\_h 1 289 8956256 3713.2  
## - source\_c 1 403 8956370 3713.2  
## - source\_t 1 1117 8957085 3713.2  
## - US 1 2328 8958296 3713.3  
## - source\_s 1 5124 8961091 3713.4  
## - source\_p 1 5765 8961733 3713.4  
## - source\_e 1 5863 8961831 3713.4  
## - source\_d 1 8028 8963996 3713.5  
## - source\_m 1 9848 8965815 3713.6  
## - source\_o 1 18480 8974447 3713.9  
## - Gender.male 1 18794 8974762 3713.9  
## - source\_b 1 20403 8976370 3714.0  
## - source\_w 1 24720 8980688 3714.2  
## - Web.order 1 26869 8982837 3714.3  
## - source\_u 1 33473 8989440 3714.5  
## - sequence\_number 1 37305 8993273 3714.7  
## - source\_a 1 42415 8998383 3714.9  
## <none> 8955968 3715.2  
## - source\_r 1 89469 9045437 3716.8  
## - Address\_is\_res 1 408448 9364416 3729.4  
## - Freq 1 4939651 13895618 3872.6  
##   
## Step: AIC=3713.18  
## Spending ~ sequence\_number + US + source\_a + source\_c + source\_b +   
## source\_d + source\_e + source\_m + source\_o + source\_h + source\_r +   
## source\_s + source\_t + source\_u + source\_p + source\_w + Freq +   
## Web.order + Gender.male + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_c 1 399 8956430 3711.2  
## - source\_h 1 575 8956606 3711.2  
## - source\_t 1 1365 8957396 3711.2  
## - US 1 2379 8958410 3711.3  
## - source\_s 1 7693 8963724 3711.5  
## - source\_p 1 7838 8963869 3711.5  
## - source\_e 1 11124 8967155 3711.6  
## - source\_d 1 13615 8969646 3711.7  
## - source\_m 1 13861 8969892 3711.7  
## - Gender.male 1 18815 8974846 3711.9  
## - source\_o 1 23418 8979449 3712.1  
## - Web.order 1 27097 8983128 3712.3  
## - source\_b 1 33502 8989533 3712.5  
## - sequence\_number 1 37254 8993285 3712.7  
## <none> 8956031 3713.2  
## - source\_w 1 51396 9007427 3713.3  
## - source\_u 1 69988 9026019 3714.0  
## - source\_a 1 91244 9047275 3714.9  
## - source\_r 1 182568 9138599 3718.5  
## - Address\_is\_res 1 408431 9364463 3727.4  
## - Freq 1 4939687 13895718 3870.6  
##   
## Step: AIC=3711.19  
## Spending ~ sequence\_number + US + source\_a + source\_b + source\_d +   
## source\_e + source\_m + source\_o + source\_h + source\_r + source\_s +   
## source\_t + source\_u + source\_p + source\_w + Freq + Web.order +   
## Gender.male + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_t 1 966 8957396 3709.2  
## - source\_h 1 1163 8957593 3709.2  
## - US 1 2409 8958840 3709.3  
## - source\_s 1 8387 8964817 3709.5  
## - source\_p 1 9935 8966365 3709.6  
## - source\_m 1 15084 8971514 3709.8  
## - source\_e 1 16523 8972953 3709.9  
## - source\_d 1 17392 8973823 3709.9  
## - Gender.male 1 18582 8975013 3709.9  
## - source\_o 1 24565 8980995 3710.2  
## - Web.order 1 26770 8983201 3710.3  
## - sequence\_number 1 37325 8993755 3710.7  
## - source\_b 1 42964 8999394 3710.9  
## <none> 8956430 3711.2  
## - source\_w 1 89847 9046278 3712.8  
## - source\_u 1 125509 9081939 3714.2  
## - source\_a 1 172175 9128605 3716.1  
## - source\_r 1 312615 9269045 3721.6  
## - Address\_is\_res 1 408213 9364643 3725.4  
## - Freq 1 4967312 13923742 3869.4  
##   
## Step: AIC=3709.23  
## Spending ~ sequence\_number + US + source\_a + source\_b + source\_d +   
## source\_e + source\_m + source\_o + source\_h + source\_r + source\_s +   
## source\_u + source\_p + source\_w + Freq + Web.order + Gender.male +   
## Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_h 1 1619 8959015 3707.3  
## - US 1 2508 8959904 3707.3  
## - source\_s 1 7448 8964844 3707.5  
## - source\_p 1 11030 8968426 3707.7  
## - source\_m 1 14125 8971521 3707.8  
## - source\_e 1 16260 8973656 3707.9  
## - source\_d 1 16617 8974013 3707.9  
## - Gender.male 1 18356 8975752 3708.0  
## - source\_o 1 23612 8981008 3708.2  
## - Web.order 1 26534 8983930 3708.3  
## - sequence\_number 1 37521 8994917 3708.7  
## - source\_b 1 42989 9000385 3709.0  
## <none> 8957396 3709.2  
## - source\_w 1 100723 9058119 3711.3  
## - source\_u 1 143294 9100690 3713.0  
## - source\_a 1 202277 9159673 3715.3  
## - source\_r 1 354714 9312110 3721.3  
## - Address\_is\_res 1 407364 9364760 3723.4  
## - Freq 1 4966347 13923743 3867.4  
##   
## Step: AIC=3707.3  
## Spending ~ sequence\_number + US + source\_a + source\_b + source\_d +   
## source\_e + source\_m + source\_o + source\_r + source\_s + source\_u +   
## source\_p + source\_w + Freq + Web.order + Gender.male + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - US 1 2779 8961794 3705.4  
## - source\_s 1 8638 8967653 3705.6  
## - source\_p 1 10518 8969533 3705.7  
## - source\_m 1 15628 8974643 3705.9  
## - Gender.male 1 18672 8977688 3706.1  
## - source\_d 1 18886 8977902 3706.1  
## - source\_e 1 19829 8978844 3706.1  
## - source\_o 1 24861 8983877 3706.3  
## - Web.order 1 27941 8986956 3706.4  
## - sequence\_number 1 37464 8996479 3706.8  
## - source\_b 1 46782 9005798 3707.2  
## <none> 8959015 3707.3  
## - source\_w 1 113333 9072348 3709.9  
## - source\_u 1 160167 9119183 3711.7  
## - source\_a 1 223935 9182950 3714.3  
## - source\_r 1 382083 9341098 3720.5  
## - Address\_is\_res 1 429216 9388231 3722.3  
## - Freq 1 4964747 13923763 3865.4  
##   
## Step: AIC=3705.41  
## Spending ~ sequence\_number + source\_a + source\_b + source\_d +   
## source\_e + source\_m + source\_o + source\_r + source\_s + source\_u +   
## source\_p + source\_w + Freq + Web.order + Gender.male + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_s 1 8984 8970778 3703.8  
## - source\_p 1 11552 8973345 3703.9  
## - source\_m 1 15336 8977129 3704.0  
## - source\_e 1 18547 8980340 3704.2  
## - source\_d 1 19660 8981454 3704.2  
## - Gender.male 1 19866 8981659 3704.2  
## - source\_o 1 25267 8987061 3704.4  
## - Web.order 1 27680 8989474 3704.5  
## - sequence\_number 1 37952 8999746 3704.9  
## - source\_b 1 46806 9008600 3705.3  
## <none> 8961794 3705.4  
## - source\_w 1 110725 9072519 3707.9  
## - source\_u 1 157432 9119226 3709.7  
## - source\_a 1 225737 9187531 3712.4  
## - source\_r 1 379664 9341458 3718.5  
## - Address\_is\_res 1 428918 9390712 3720.4  
## - Freq 1 4963964 13925758 3863.4  
##   
## Step: AIC=3703.77  
## Spending ~ sequence\_number + source\_a + source\_b + source\_d +   
## source\_e + source\_m + source\_o + source\_r + source\_u + source\_p +   
## source\_w + Freq + Web.order + Gender.male + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_m 1 11792 8982570 3702.3  
## - source\_e 1 11972 8982750 3702.3  
## - source\_p 1 13933 8984711 3702.3  
## - source\_d 1 14309 8985087 3702.4  
## - Gender.male 1 19856 8990634 3702.6  
## - source\_o 1 21835 8992613 3702.7  
## - Web.order 1 28945 8999723 3702.9  
## - sequence\_number 1 37769 9008547 3703.3  
## - source\_b 1 40019 9010797 3703.4  
## <none> 8970778 3703.8  
## - source\_w 1 102658 9073436 3705.9  
## - source\_u 1 151284 9122062 3707.8  
## - source\_a 1 224855 9195633 3710.8  
## - source\_r 1 382525 9353303 3716.9  
## - Address\_is\_res 1 425445 9396223 3718.6  
## - Freq 1 4958105 13928883 3861.5  
##   
## Step: AIC=3702.25  
## Spending ~ sequence\_number + source\_a + source\_b + source\_d +   
## source\_e + source\_o + source\_r + source\_u + source\_p + source\_w +   
## Freq + Web.order + Gender.male + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_e 1 7218 8989788 3700.5  
## - source\_d 1 10358 8992929 3700.7  
## - source\_p 1 15616 8998186 3700.9  
## - source\_o 1 19194 9001764 3701.0  
## - Gender.male 1 19241 9001811 3701.0  
## - Web.order 1 30427 9012997 3701.5  
## - source\_b 1 34445 9017015 3701.6  
## - sequence\_number 1 35193 9017764 3701.7  
## <none> 8982570 3702.3  
## - source\_w 1 91622 9074192 3703.9  
## - source\_u 1 139630 9122200 3705.9  
## - source\_a 1 213223 9195794 3708.8  
## - source\_r 1 371531 9354101 3715.0  
## - Address\_is\_res 1 417272 9399842 3716.7  
## - Freq 1 4956392 13938962 3859.8  
##   
## Step: AIC=3700.54  
## Spending ~ sequence\_number + source\_a + source\_b + source\_d +   
## source\_o + source\_r + source\_u + source\_p + source\_w + Freq +   
## Web.order + Gender.male + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_d 1 5850 8995638 3698.8  
## - source\_o 1 15927 9005715 3699.2  
## - Gender.male 1 18330 9008118 3699.3  
## - source\_p 1 18991 9008779 3699.3  
## - source\_b 1 28210 9017998 3699.7  
## - Web.order 1 31470 9021258 3699.8  
## - sequence\_number 1 34618 9024407 3699.9  
## <none> 8989788 3700.5  
## - source\_w 1 91519 9081307 3702.2  
## - source\_u 1 147782 9137570 3704.5  
## - source\_a 1 243448 9233237 3708.2  
## - source\_r 1 406378 9396166 3714.6  
## - Address\_is\_res 1 422506 9412294 3715.2  
## - Freq 1 4990447 13980236 3858.8  
##   
## Step: AIC=3698.78  
## Spending ~ sequence\_number + source\_a + source\_b + source\_o +   
## source\_r + source\_u + source\_p + source\_w + Freq + Web.order +   
## Gender.male + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_o 1 14786 9010424 3697.4  
## - Gender.male 1 17444 9013083 3697.5  
## - source\_p 1 20037 9015675 3697.6  
## - source\_b 1 25737 9021375 3697.8  
## - Web.order 1 31040 9026679 3698.0  
## - sequence\_number 1 32211 9027849 3698.1  
## <none> 8995638 3698.8  
## - source\_w 1 85738 9081377 3700.2  
## - source\_u 1 141999 9137638 3702.5  
## - source\_a 1 239275 9234913 3706.3  
## - source\_r 1 402302 9397941 3712.7  
## - Address\_is\_res 1 428660 9424298 3713.7  
## - Freq 1 5022153 14017792 3857.8  
##   
## Step: AIC=3697.37  
## Spending ~ sequence\_number + source\_a + source\_b + source\_r +   
## source\_u + source\_p + source\_w + Freq + Web.order + Gender.male +   
## Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - Gender.male 1 16232 9026656 3696.0  
## - source\_p 1 20483 9030907 3696.2  
## - source\_b 1 24006 9034430 3696.3  
## - sequence\_number 1 30831 9041255 3696.6  
## - Web.order 1 31903 9042327 3696.7  
## <none> 9010424 3697.4  
## - source\_w 1 79977 9090401 3698.6  
## - source\_u 1 135060 9145484 3700.8  
## - source\_a 1 230845 9241269 3704.6  
## - source\_r 1 393041 9403465 3710.9  
## - Address\_is\_res 1 425031 9435455 3712.1  
## - Freq 1 5007370 14017794 3855.8  
##   
## Step: AIC=3696.03  
## Spending ~ sequence\_number + source\_a + source\_b + source\_r +   
## source\_u + source\_p + source\_w + Freq + Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_p 1 18242 9044898 3694.8  
## - source\_b 1 23104 9049760 3695.0  
## - sequence\_number 1 28993 9055648 3695.2  
## - Web.order 1 31537 9058193 3695.3  
## <none> 9026656 3696.0  
## - source\_w 1 82789 9109445 3697.3  
## - source\_u 1 132142 9158798 3699.3  
## - source\_a 1 232248 9258904 3703.2  
## - source\_r 1 387734 9414390 3709.3  
## - Address\_is\_res 1 450294 9476950 3711.7  
## - Freq 1 4998473 14025128 3854.0  
##   
## Step: AIC=3694.76  
## Spending ~ sequence\_number + source\_a + source\_b + source\_r +   
## source\_u + source\_w + Freq + Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - source\_b 1 24557 9069454 3693.7  
## - sequence\_number 1 26848 9071746 3693.8  
## - Web.order 1 31811 9076709 3694.0  
## <none> 9044898 3694.8  
## - source\_w 1 88683 9133580 3696.3  
## - source\_u 1 139980 9184877 3698.3  
## - source\_a 1 244763 9289661 3702.5  
## - source\_r 1 398358 9443256 3708.4  
## - Address\_is\_res 1 442754 9487652 3710.1  
## - Freq 1 4984457 14029354 3852.1  
##   
## Step: AIC=3693.75  
## Spending ~ sequence\_number + source\_a + source\_r + source\_u +   
## source\_w + Freq + Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - sequence\_number 1 28190 9097644 3692.9  
## - Web.order 1 31843 9101298 3693.0  
## <none> 9069454 3693.7  
## - source\_w 1 76449 9145903 3694.8  
## - source\_u 1 125241 9194696 3696.7  
## - source\_a 1 226002 9295457 3700.7  
## - source\_r 1 379750 9449205 3706.6  
## - Address\_is\_res 1 432648 9502102 3708.7  
## - Freq 1 4974600 14044054 3850.5  
##   
## Step: AIC=3692.87  
## Spending ~ source\_a + source\_r + source\_u + source\_w + Freq +   
## Web.order + Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## - Web.order 1 28468 9126112 3692.0  
## <none> 9097644 3692.9  
## - source\_w 1 72513 9170157 3693.8  
## - source\_u 1 128409 9226053 3696.0  
## - source\_a 1 235160 9332804 3700.1  
## - source\_r 1 374471 9472115 3705.5  
## - Address\_is\_res 1 437501 9535146 3707.9  
## - Freq 1 4970459 14068103 3849.1  
##   
## Step: AIC=3692.01  
## Spending ~ source\_a + source\_r + source\_u + source\_w + Freq +   
## Address\_is\_res  
##   
## Df Sum of Sq RSS AIC  
## <none> 9126112 3692.0  
## - source\_w 1 68461 9194574 3692.7  
## - source\_u 1 118458 9244570 3694.7  
## - source\_a 1 218222 9344334 3698.6  
## - source\_r 1 364657 9490769 3704.2  
## - Address\_is\_res 1 441738 9567851 3707.2  
## - Freq 1 4971763 14097876 3847.9

# For the Presentation  
plot(lm\_model)



1. Regression trees

library(rpart)  
tree\_model <- rpart(Spending ~ ., data = train.data)  
tree\_model

## n= 396   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 396 20852660.0 218.1288   
## 2) Freq< 7.5 389 13884290.0 201.5219   
## 4) Freq< 2.5 279 3633912.0 147.5125 \*  
## 5) Freq>=2.5 110 7372329.0 338.5091   
## 10) sequence\_number>=493.5 86 4034302.0 312.3488   
## 20) Freq< 3.5 46 2063335.0 244.1957 \*  
## 21) Freq>=3.5 40 1511590.0 390.7250 \*  
## 11) sequence\_number< 493.5 24 3068276.0 432.2500   
## 22) sequence\_number< 362 15 373282.9 305.7333 \*  
## 23) sequence\_number>=362 9 2054735.0 643.1111 \*  
## 3) Freq>=7.5 7 899246.0 1141.0000 \*

summary(tree\_model)

## Call:  
## rpart(formula = Spending ~ ., data = train.data)  
## n= 396   
##   
## CP nsplit rel error xerror xstd  
## 1 0.29104785 0 1.0000000 1.0036310 0.1905959  
## 2 0.13801837 1 0.7089522 0.9749312 0.1799703  
## 3 0.02188988 2 0.5709338 0.6932199 0.1240769  
## 4 0.01000000 5 0.5052642 0.7440764 0.1353690  
##   
## Variable importance  
## Freq sequence\_number source\_r source\_h Web.order   
## 87 9 1 1 1   
##   
## Node number 1: 396 observations, complexity param=0.2910478  
## mean=218.1288, MSE=52658.23   
## left son=2 (389 obs) right son=3 (7 obs)  
## Primary splits:  
## Freq < 7.5 to the left, improve=0.291047800, (0 missing)  
## sequence\_number < 831.5 to the left, improve=0.008679098, (0 missing)  
## source\_s < 0.5 to the right, improve=0.006414448, (0 missing)  
## Web.order < 0.5 to the right, improve=0.006057373, (0 missing)  
## US < 0.5 to the right, improve=0.002103188, (0 missing)  
##   
## Node number 2: 389 observations, complexity param=0.1380184  
## mean=201.5219, MSE=35692.26   
## left son=4 (279 obs) right son=5 (110 obs)  
## Primary splits:  
## Freq < 2.5 to the left, improve=0.207288200, (0 missing)  
## source\_s < 0.5 to the right, improve=0.006397333, (0 missing)  
## sequence\_number < 1770 to the right, improve=0.004004395, (0 missing)  
## source\_w < 0.5 to the left, improve=0.002714116, (0 missing)  
## Web.order < 0.5 to the right, improve=0.002701985, (0 missing)  
## Surrogate splits:  
## source\_h < 0.5 to the left, agree=0.725, adj=0.027, (0 split)  
##   
## Node number 3: 7 observations  
## mean=1141, MSE=128463.7   
##   
## Node number 4: 279 observations  
## mean=147.5125, MSE=13024.77   
##   
## Node number 5: 110 observations, complexity param=0.02188988  
## mean=338.5091, MSE=67021.18   
## left son=10 (86 obs) right son=11 (24 obs)  
## Primary splits:  
## sequence\_number < 493.5 to the right, improve=0.036589720, (0 missing)  
## Freq < 4.5 to the left, improve=0.035030410, (0 missing)  
## Address\_is\_res < 0.5 to the right, improve=0.027501800, (0 missing)  
## US < 0.5 to the right, improve=0.023158160, (0 missing)  
## source\_b < 0.5 to the right, improve=0.005859253, (0 missing)  
## Surrogate splits:  
## source\_s < 0.5 to the left, agree=0.791, adj=0.042, (0 split)  
##   
## Node number 10: 86 observations, complexity param=0.02188988  
## mean=312.3488, MSE=46910.48   
## left son=20 (46 obs) right son=21 (40 obs)  
## Primary splits:  
## Freq < 3.5 to the left, improve=0.113867600, (0 missing)  
## sequence\_number < 853 to the left, improve=0.046666030, (0 missing)  
## Address\_is\_res < 0.5 to the right, improve=0.030608160, (0 missing)  
## US < 0.5 to the right, improve=0.017884590, (0 missing)  
## Web.order < 0.5 to the left, improve=0.002861171, (0 missing)  
## Surrogate splits:  
## source\_c < 0.5 to the left, agree=0.581, adj=0.100, (0 split)  
## sequence\_number < 1331.5 to the right, agree=0.570, adj=0.075, (0 split)  
## source\_o < 0.5 to the left, agree=0.558, adj=0.050, (0 split)  
## source\_x < 0.5 to the left, agree=0.558, adj=0.050, (0 split)  
## Gender.male < 0.5 to the left, agree=0.558, adj=0.050, (0 split)  
##   
## Node number 11: 24 observations, complexity param=0.02188988  
## mean=432.25, MSE=127844.9   
## left son=22 (15 obs) right son=23 (9 obs)  
## Primary splits:  
## sequence\_number < 362 to the left, improve=0.2086705000, (0 missing)  
## Web.order < 0.5 to the right, improve=0.0466250500, (0 missing)  
## Address\_is\_res < 0.5 to the right, improve=0.0375220600, (0 missing)  
## Freq < 3.5 to the right, improve=0.0027020380, (0 missing)  
## source\_u < 0.5 to the left, improve=0.0007845356, (0 missing)  
## Surrogate splits:  
## source\_r < 0.5 to the left, agree=0.708, adj=0.222, (0 split)  
## Web.order < 0.5 to the right, agree=0.667, adj=0.111, (0 split)  
##   
## Node number 20: 46 observations  
## mean=244.1957, MSE=44855.11   
##   
## Node number 21: 40 observations  
## mean=390.725, MSE=37789.75   
##   
## Node number 22: 15 observations  
## mean=305.7333, MSE=24885.53   
##   
## Node number 23: 9 observations  
## mean=643.1111, MSE=228303.9

tree\_model <- rpart(Spending ~ ., data = valid.data)  
tree\_model

## n= 363   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 363 14647510.0 190.7052   
## 2) Freq< 4.5 345 8277597.0 168.2783   
## 4) Freq< 1.5 186 3018052.0 133.6183 \*  
## 5) Freq>=1.5 159 4774713.0 208.8239   
## 10) source\_r< 0.5 146 2792097.0 196.9795   
## 20) Freq< 2.5 96 1768177.0 170.0104 \*  
## 21) Freq>=2.5 50 820035.1 248.7600 \*  
## 11) source\_r>=0.5 13 1732100.0 341.8462 \*  
## 3) Freq>=4.5 18 2870500.0 620.5556 \*

summary(tree\_model)

## Call:  
## rpart(formula = Spending ~ ., data = valid.data)  
## n= 363   
##   
## CP nsplit rel error xerror xstd  
## 1 0.23890810 0 1.0000000 1.0029386 0.2308859  
## 2 0.03309999 1 0.7610919 0.8534746 0.2008220  
## 3 0.01710301 2 0.7279919 0.8443229 0.1999892  
## 4 0.01391942 3 0.7108889 0.8561694 0.2008783  
## 5 0.01000000 4 0.6969695 0.8177161 0.1773704  
##   
## Variable importance  
## Freq source\_r Address\_is\_res US   
## 91 5 1 1   
##   
## Node number 1: 363 observations, complexity param=0.2389081  
## mean=190.7052, MSE=40351.25   
## left son=2 (345 obs) right son=3 (18 obs)  
## Primary splits:  
## Freq < 4.5 to the left, improve=0.238908100, (0 missing)  
## sequence\_number < 39 to the right, improve=0.033091490, (0 missing)  
## source\_a < 0.5 to the left, improve=0.012874810, (0 missing)  
## source\_r < 0.5 to the left, improve=0.010148240, (0 missing)  
## Address\_is\_res < 0.5 to the right, improve=0.008866357, (0 missing)  
##   
## Node number 2: 345 observations, complexity param=0.03309999  
## mean=168.2783, MSE=23993.04   
## left son=4 (186 obs) right son=5 (159 obs)  
## Primary splits:  
## Freq < 1.5 to the left, improve=0.058571620, (0 missing)  
## source\_r < 0.5 to the left, improve=0.034469180, (0 missing)  
## sequence\_number < 373.5 to the right, improve=0.018573100, (0 missing)  
## Address\_is\_res < 0.5 to the right, improve=0.008751507, (0 missing)  
## source\_e < 0.5 to the right, improve=0.008304255, (0 missing)  
## Surrogate splits:  
## Address\_is\_res < 0.5 to the left, agree=0.597, adj=0.126, (0 split)  
## US < 0.5 to the right, agree=0.568, adj=0.063, (0 split)  
## sequence\_number < 1757 to the left, agree=0.557, adj=0.038, (0 split)  
## source\_b < 0.5 to the left, agree=0.548, adj=0.019, (0 split)  
## source\_h < 0.5 to the left, agree=0.548, adj=0.019, (0 split)  
##   
## Node number 3: 18 observations  
## mean=620.5556, MSE=159472.2   
##   
## Node number 4: 186 observations  
## mean=133.6183, MSE=16226.09   
##   
## Node number 5: 159 observations, complexity param=0.01710301  
## mean=208.8239, MSE=30029.64   
## left son=10 (146 obs) right son=11 (13 obs)  
## Primary splits:  
## source\_r < 0.5 to the left, improve=0.05246733, (0 missing)  
## Address\_is\_res < 0.5 to the right, improve=0.05020628, (0 missing)  
## Freq < 2.5 to the left, improve=0.03085597, (0 missing)  
## sequence\_number < 390.5 to the right, improve=0.02508624, (0 missing)  
## Gender.male < 0.5 to the left, improve=0.01955449, (0 missing)  
##   
## Node number 10: 146 observations, complexity param=0.01391942  
## mean=196.9795, MSE=19123.95   
## left son=20 (96 obs) right son=21 (50 obs)  
## Primary splits:  
## Freq < 2.5 to the left, improve=0.07302212, (0 missing)  
## Address\_is\_res < 0.5 to the right, improve=0.05922558, (0 missing)  
## source\_a < 0.5 to the left, improve=0.02225098, (0 missing)  
## source\_u < 0.5 to the left, improve=0.02127196, (0 missing)  
## sequence\_number < 98 to the right, improve=0.01698261, (0 missing)  
## Surrogate splits:  
## source\_t < 0.5 to the left, agree=0.671, adj=0.04, (0 split)  
## source\_h < 0.5 to the left, agree=0.664, adj=0.02, (0 split)  
##   
## Node number 11: 13 observations  
## mean=341.8462, MSE=133238.4   
##   
## Node number 20: 96 observations  
## mean=170.0104, MSE=18418.51   
##   
## Node number 21: 50 observations  
## mean=248.76, MSE=16400.7

1. Choose one model on the basis of its performance on the validation data.

# make predictions using multiple linear  
lm\_predictions <- predict(lm\_model, valid.data)  
# Calculate RMSE  
lm\_rmse <- sqrt(mean((valid.data$Spending - lm\_predictions) ^ 2))  
# make predictions using regression tree  
tree\_predictions <- predict(tree\_model, valid.data)  
#calculate RMSE   
tree\_rmse <- sqrt(mean((valid.data$Spending - tree\_predictions) ^ 2))

if (lm\_rmse < tree\_rmse) { "lm" } else{"tree"}

## [1] "lm"

# if (lm\_rmse < tree\_rmse) {chosen\_model <- lm\_model} else{chosen\_model <- tree\_model}  
# chosen\_model

1. Return to the original test data partition. Note that this test data partition includes both purchasers and non-purchasers. Create a new data frame called Score Analysis that contains the test data portion of this data-set.

score\_analysis <- test.data

1. Add a column to the data frame with the predicted scores from the logistic regression.

score\_analysis$lr\_prediction <- predict(lr\_model\_b, newdata = score\_analysis)

1. Add another column with the predicted spending amount from he prediction model chosen.

score\_analysis$lm\_prediction <- predict(lm\_model, newdata = score\_analysis)

1. Add a column for “adjusted probability of purchase” by multiplying “predicted probability of purchase” by 0.107. This is to adjust for oversampling the purchasers (see earlier description).

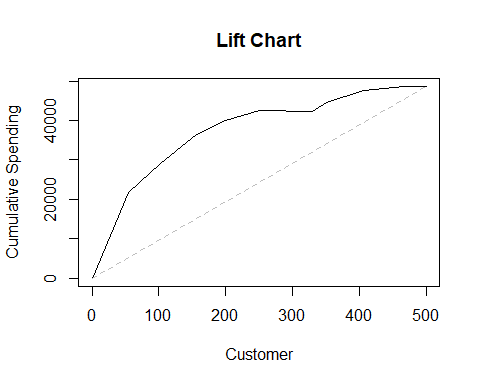
score\_analysis$adjusted\_probability\_purchase <-score\_analysis$lr\_prediction \* 0.107

1. Add a column for expected spending: adjusted probability of purchase × predicted spending.

score\_analysis$expected\_spending <-score\_analysis$adjusted\_probability\_purchase \* score\_analysis$lm\_prediction

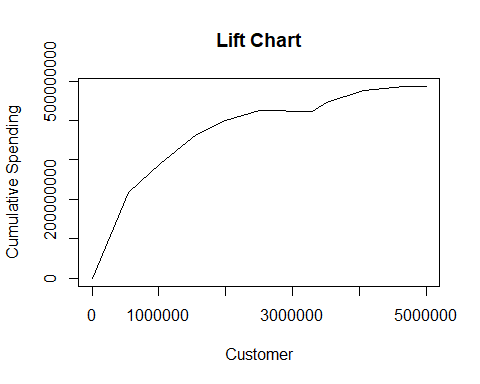
1. Plot the lift chart of the expected spending.

Spending <- score\_analysis$Spending-2  
exp\_spending <-score\_analysis$expected\_spending-2  
  
library(gains)  
gain <- gains(Spending, exp\_spending)# predictions  
# cumulative lift chart  
options(scipen=999) # avoid scientific notation  
# we will compute the gain relative to price  
  
plot(c(0,gain$cume.pct.of.total\*sum(Spending))~c(0,gain$cume.obs),  
xlab="Customer", ylab="Cumulative Spending", main="Lift Chart", type="l")  
# baseline  
lines(c(0,sum(Spending))~c(0,dim(score\_analysis)[1]), col="gray", lty=2)



1. Using this lift curve, estimate the gross profit that would result from mailing to the 180,000 names on the basis of your data mining models.

factor <- 5000000/500   
  
plot(c(0,factor \* gain$cume.pct.of.total\*sum(Spending))~c(0,factor\*gain$cume.obs),  
xlab="Customer", ylab="Cumulative Spending", main="Lift Chart", type="l")



final\_gross\_profit <- approx(c(0,factor\*gain$cume.obs), c(0,factor \* gain$cume.pct.of.total\*sum(Spending)), xout=180000)  
final\_gross\_profit

## $x  
## [1] 180000  
##   
## $y  
## [1] 72186667