RSM 8522: Assignment 4

Erin (Kexin) Zhu - 1008241751

3/12/2023

Part I: Calibrating models using the Estimation Sample

Question 1

Estimate a logistic regression model using "Choice" as the dependent variable and the following as explanatory variables: Gender, Amt_purchased, Frequency, Last_Purchase, First_purchase, P_Child, P_Youth, P_Cook, P_DIY, and P_Art. Report the regression results.

```
data_est <- read.csv("~/Downloads/estimationsample.csv")
model <- glm(Choice ~ Gender + Amt_purchased + Frequency + Last_Purchase + First_purchase + P_Child + P_
summary(model)
```

```
##
## Call:
## glm(formula = Choice ~ Gender + Amt_purchased + Frequency + Last_Purchase +
##     First_purchase + P_Child + P_Youth + P_Cook + P_DIY + P_Art,
##     family = binomial(), data = data_est)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8096  -0.4261  -0.2882  -0.1789   2.8546
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.4637596  0.3675407  -3.983 6.82e-05 ***
## Gender         -0.6203510  0.2427137  -2.556 0.010591 *
## Amt_purchased   0.0013480  0.0013564   0.994 0.320329
## Frequency      -0.1098280  0.0286972  -3.827 0.000130 ***
## Last_Purchase   0.5556901  0.1614560   3.442 0.000578 ***
## First_purchase  0.0003009  0.0206859   0.015 0.988395
## P_Child        -0.8590171  0.2065449  -4.159 3.20e-05 ***
## P_Youth        -0.3827273  0.2384632  -1.605 0.108499
## P_Cook         -0.6094687  0.1935532  -3.149 0.001639 **
## P_DIY          -1.1577349  0.2543039  -4.553 5.30e-06 ***
## P_Art           0.5209002  0.2131452   2.444 0.014530 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 660.94  on 1099  degrees of freedom
## Residual deviance: 536.94  on 1089  degrees of freedom
## AIC: 558.94
##
## Number of Fisher Scoring iterations: 6
```

```
model %>%
  broom::tidy() %>%
  knitr::kable(caption = 'Logistic Regression Model')
```

Table 1: Logistic Regression Model

term	estimate	std.error	statistic	p.value
(Intercept)	-1.4637596	0.3675407	-3.9825780	0.0000682
Gender	-0.6203510	0.2427137	-2.5558961	0.0105915
Amt_purchased	0.0013480	0.0013564	0.9937822	0.3203289
Frequency	-0.1098280	0.0286972	-3.8271375	0.0001296
Last_Purchase	0.5556901	0.1614560	3.4417426	0.0005780
First_purchase	0.0003009	0.0206859	0.0145452	0.9883950
P_Child	-0.8590171	0.2065449	-4.1589853	0.0000320
P_Youth	-0.3827273	0.2384632	-1.6049744	0.1084994
P_Cook	-0.6094687	0.1935532	-3.1488433	0.0016392
P_DIY	-1.1577349	0.2543039	-4.5525652	0.0000053
P_Art	0.5209002	0.2131452	2.4438747	0.0145305

Question 2

Summarize and interpret the results (so that a marketing manager can understand them). Which variables are statistically significant? Which seem to be economically "important"? Interpret the economic importance for some of the explanatory variables.

According to the model output in Question 1, variables *Gender*, *Frequency*, *Last_Purchase*, *P_Child*, *P_Cook*, *P_DIY*, and *P_Art* are statistically significant at the 0.05 significance level. Variables *Amt_purchased*, *First_purchase* and *P_Youth* are statistically insignificant.

To evaluate the economical importance of the variables, the statistically insignificant variables are excluded from further analysis. By estimating the marginal probability change for each variable, it can be concluded that variables *P_DIY*, *P_Child* and *Gender* have the biggest "economical" importance. With all other variables held constant,one unit of increase in the number of DIY books decreases the probability of purchasing by 7.9%. Similarly, one unit of increase in the number of Child books decreases the probability of purchasing by 5.9%. With all other variables held constant, males decreases the probability of purchasing by 4.6% compared to females.

Another insight is that variables *P_Art* and *Last_Purchase* have positive impacts on the probability of purchasing while all other variables have negative impacts. With all other variables held constant, one unit of increase in the number of Art books increases the probability of purchasing by 3.6%. One unit of increase in the number of Last_Purchases increases the probability of purchasing by 2.06e-5.

```
library("mfx")
mfx <- logitmfx(Choice ~ Gender + Amt_purchased + Frequency + Last_Purchase + First_purchase + P_Child +
mfx$mfxest
```

```
##                       dF/dx    Std. Err.           z        P>|z|
## Gender         -4.569518e-02 1.898176e-02 -2.40731995 1.607008e-02
## Amt_purchased   9.252661e-05 9.368492e-05  0.98763617 3.233309e-01
## Frequency      -7.538627e-03 2.142570e-03 -3.51849704 4.339987e-04
## Last_Purchase   3.814274e-02 1.196509e-02  3.18783556 1.433420e-03
## First_purchase  2.065252e-05 1.419904e-03  0.01454501 9.883952e-01
## P_Child        -5.896319e-02 1.579999e-02 -3.73185040 1.900784e-04
## P_Youth        -2.627052e-02 1.667689e-02 -1.57526487 1.151953e-01
## P_Cook         -4.183412e-02 1.417704e-02 -2.95083610 3.169150e-03
## P_DIY          -7.946727e-02 1.975143e-02 -4.02336785 5.737178e-05
## P_Art           3.575475e-02 1.518504e-02  2.35460363 1.854247e-02
```

Question 3

Run a "pure R" model (i.e., a RFM model without F and M) with deciles and predict response rate for each decile group.

Below is the table that shows the Recency quintiles and their corresponding response rats. Quintiles are reordered so that the smaller quintiles have higher response rates.

```
R_quintiles <- .bincode(data_est$Last_Purchase, quantile(data_est$Last_Purchase, probs = seq(0, 1, 0.1)
data_est$R_quintiles <- R_quintiles
# Reorder
data_est$R_quintiles <- (11 - data_est$R_quintiles)
library(dplyr)
data_est %>%
  group_by(R_quintiles) %>%
  summarize(response_rate = mean(Choice))
```

```
## # A tibble: 5 x 2
##   R_quintiles response_rate
##         <dbl>         <dbl>
## 1           1        0.173
## 2           2        0.114
## 3           3        0.102
## 4           7        0.0749
## 5          10        0.0705
```

Part II: Validate the two models on the Validation Sample

Question 4

Check the overall classification performance of the logistic regression model above on the validation sample. Create a table showing the fraction of observations which are correctly predicted by the model. (Hint: use table () command shown in the lecture slides.)

According to the matrix output below, the model accuracy is 91.3%.

```
data_val <- read.csv("~/Downloads/validation1.csv")
data_val<- subset(data_val, select = 1:12)
pred <- predict(model, newdata = data_val, type = "response")
pred_class <- ifelse(pred >= 0.5, 1, 0)
```

```r
conf_mat <- table(data_val$Choice, pred_class)
accuracy <- sum(diag(conf_mat)) / sum(conf_mat)
cat("Overall classification accuracy:", round(accuracy, 3))
```

```
## Overall classification accuracy: 0.913
```

```r
conf_mat %>%
  knitr::kable(caption = 'Confusion Matrix')
```

Table 2: Confusion Matrix

	0	1
0	1087	12
1	92	9

Question 5

Using your logistic regression result, assign each customer to a decile based on his/her predicted probability of purchase. (Hint: use .bincode function as we did in RFM.)

A new column *probs* is created and added to the validation set that shows the predicted probability of purchasing. Corresponding code is shown as below. Each customer is assigned to a decile based on his/her predicted probability of purchase. The deciles are reverted so that the smaller deciles have higher probabilities.

```r
data_val$probs <- predict(model, newdata = data_val, type = "response")
# Bin predicted probabilities into deciles
data_val$prob_decile <- .bincode(data_val$probs, quantile(data_val$probs, probs = seq(0, 1, 0.1),na.rm =
# Reorder
data_val$prob_decile <- (11 - data_val$prob_decile)
head(data_val, 10)
```

```
##    Intercept Gender Amt_purchased Frequency Last_Purchase First_purchase
## 1          1      0            42        18             2             24
## 2          1      1           163         4             1              4
## 3          1      1           228        10            10             64
## 4          1      0           263         8             2             14
## 5          1      1           252        12             2             18
## 6          1      0            92        16             1             16
## 7          1      1           220         2             1              2
## 8          1      0           310        12             2             14
## 9          1      0           288        16             1             16
## 10         1      0           151        32             1             32
##    P_Child P_Youth P_Cook P_DIY P_Art Choice       probs prob_decile
## 1        1       0      1     0     0      0 0.023340863           8
## 2        0       0      0     0     1      0 0.226898804           1
## 3        2       2      3     1     0      0 0.059049176           5
## 4        0       0      1     0     0      0 0.185165111           2
## 5        0       0      0     1     1      0 0.070280352           4
## 6        0       0      0     0     0      0 0.073339502           4
```

```
## 7        0       0      0     1     0      0 0.068583719           4
## 8        0       1      0     0     1      0 0.247861137           1
## 9        0       0      1     0     0      0 0.053063131           5
## 10       0       0      1     0     0      0 0.008011299          10
```

Question 6

Using your logistic regression result, report the number of customers, the number of buyers of "The Art History of Florence," and the average response rate to the offer by decile for the 1200 customers in the validation dataset, respectively.

```
# Number of customers by decile
num_customers<-table(data_val$prob_decile)
num_customers
```

```
##
##   1   2   3   4   5   6   7   8   9  10
## 120 120 120 120 120 120 120 120 120 120
```

```
# Number of buyers by decile
num_buyers<-tapply(data_val$Choice, data_val$prob_decile, sum)
num_buyers
```

```
##  1  2  3  4  5  6  7  8  9 10
## 29 23  9  9  6  7  8  6  3  1
```

```
# Average predicted response rate by decile
avg_response<-tapply(data_val$probs, data_val$prob_decile, mean)
avg_response
```

```
##           1           2           3           4           5           6
## 0.351406528 0.166445999 0.106694166 0.077520612 0.059417051 0.045362861
##           7           8           9          10
## 0.034186239 0.025072845 0.015981607 0.006208947
```

```
actual_response <-num_buyers/num_customers
actual_response
```

```
##           1           2           3           4           5           6
## 0.241666667 0.191666667 0.075000000 0.075000000 0.050000000 0.058333333
##           7           8           9          10
## 0.066666667 0.050000000 0.025000000 0.008333333
```

```
# Combine results into data frame
results <- data.frame(Decile = 1:10, Num_Customers = num_customers, Num_Buyers = num_buyers, Avg_Respons
results
```

```
##    Decile Num_Customers.Var1 Num_Customers.Freq Num_Buyers Avg_Response_Rate
## 1       1                  1                120         29       0.351406528
## 2       2                  2                120         23       0.166445999
```

```
## 3       3                  3                 120          9       0.106694166
## 4       4                  4                 120          9       0.077520612
## 5       5                  5                 120          6       0.059417051
## 6       6                  6                 120          7       0.045362861
## 7       7                  7                 120          8       0.034186239
## 8       8                  8                 120          6       0.025072845
## 9       9                  9                 120          3       0.015981607
## 10     10                 10                 120          1       0.006208947
##    Actual_Reponse.Var1 Actual_Reponse.Freq
## 1                    1         0.241666667
## 2                    2         0.191666667
## 3                    3         0.075000000
## 4                    4         0.075000000
## 5                    5         0.050000000
## 6                    6         0.058333333
## 7                    7         0.066666667
## 8                    8         0.050000000
## 9                    9         0.025000000
## 10                  10         0.008333333
```

Question 7

Using your pure-R model, report the number of customers and the number of buyers of "The Art History of Florence," by decile for the 1200 customers in the validation dataset.

The table below shows that number of customers and the number of buyers of "The Art History of Florence," by decile for the 1200 customers. In addition, the table also includes the response rate for each R quintile.

```
R_quintiles_val <- .bincode(data_val$Last_Purchase, quantile(data_est$Last_Purchase, probs = seq(0, 1, 
data_val$R_quintiles <- R_quintiles_val
# Reorder
data_val$R_quintiles <- (11 - data_val$R_quintiles)

florence_buyers <- data_val %>% filter(Choice == 1)

book_data <- data_val %>%
  group_by(R_quintiles) %>%
  summarize(num_customers = n())

book_data <- left_join(book_data, florence_buyers %>% group_by(R_quintiles) %>% summarize(num_buyers = 

# Calculate the response rate for each R quintile
book_data$Avg_Response_Rate <- book_data$num_buyers / book_data$num_customers
```

```
book_data %>%
  knitr::kable(caption = "Pure-R Model: Number of buyers in each recency decile")
```

Table 3: Pure-R Model: Number of buyers in each recency decile

R_quintiles	num_customers	num_buyers	Avg_Response_Rate
1	148	13	0.0878378
2	75	10	0.1333333

R_quintiles	num_customers	num_buyers	Avg_Response_Rate
3	152	18	0.1184211
7	386	28	0.0725389
10	439	32	0.0728929
NA	2	NA	NA

Part III: Lift and Cumulative Lift in the Validation Sample

Question 8

Use the computations above to create a table showing the lift and cumulative lift for each decile, for both logistic regression results and R(FM) results. You may want to use Excel for these calculations.

Refer to *Figure 1*.

RFM Model									
Decile	Customers	Cum Custom	Cum % Customers	Buyer	Cum Buyer	Response Rate	Lift	Cum Res Rate	Cum Lift
1	148	148	12.3%	13	13	8.8%	104.4	8.8%	104.4
2	75	223	18.6%	10	23	13.3%	158.4	10.3%	122.5
3	152	375	31.3%	18	41	11.8%	140.7	10.9%	129.9
7	386	761	63.4%	28	69	7.3%	86.2	9.1%	107.7
10	439	1200	100.0%	32	101	7.3%	86.6	8.4%	100.0
	1200			101		8.4%			
Logistics Model									
Decile	Customers	Cum Custom	Cum % Customers	Buyers	Cum Buyers	Response Rate	Lift	Cum Res Rate	Cum Lift
1	120	120	10.0%	29	29	24.2%	287.1	24.2%	287.1
2	120	240	20.0%	23	52	19.2%	227.7	21.7%	257.4
3	120	360	30.0%	9	61	7.5%	89.1	16.9%	201.3
4	120	480	40.0%	9	70	7.5%	89.1	14.6%	173.3
5	120	600	50.0%	6	76	5.0%	59.4	12.7%	150.5
6	120	720	60.0%	7	83	5.8%	69.3	11.5%	137.0
7	120	840	70.0%	8	91	6.7%	79.2	10.8%	128.7
8	120	960	80.0%	6	97	5.0%	59.4	10.1%	120.0
9	120	1080	90.0%	3	100	2.5%	29.7	9.3%	110.0
10	120	1200	100.0%	1	101	0.8%	9.9	8.4%	100.0
	1200			101		8.4%			

Figure 1: Lift and cumulative lift for RFM and logistic regression

Question 9

Use the computations above to create a table showing the gains and cumulative gains for each decile, for both logistic regression results and R(FM) results. You may want to use Excel for these calculations.

Refer to *Figure 2*.

Question 10

Create a chart showing the cumulative gains by decile along with a reference line corresponding to "no model," for the logistic regression and R(FM).

RFM Model									
Decile	Customers	Cum Custom	Cum % Customers	Buyer	Cum Buyer	Response Rate	Cum Gains	Reference Line	
1	148	148	12.3%	13	13	8.8%	12.9%	12.3%	
2	75	223	18.6%	10	23	13.3%	22.8%	18.6%	
3	152	375	31.3%	18	41	11.8%	40.6%	31.3%	
7	386	761	63.4%	28	69	7.3%	68.3%	63.4%	
10	439	1200	100.0%	32	101	7.3%	100.0%	100.0%	
	1200			101					
Logistics Model									
Decile	Customers	Cum Custome	Cum % Customers	Buyers	Cum Buyers	Response Rate	Gains	Cum Gains	Reference Line
1	120	120	10.0%	29	29	24.2%	28.7%	28.7%	10.0%
2	120	240	20.0%	23	52	19.2%	22.8%	51.5%	20.0%
3	120	360	30.0%	9	61	7.5%	8.9%	60.4%	30.0%
4	120	480	40.0%	9	70	7.5%	8.9%	69.3%	40.0%
5	120	600	50.0%	6	76	5.0%	5.9%	75.2%	50.0%
6	120	720	60.0%	7	83	5.8%	6.9%	82.2%	60.0%
7	120	840	70.0%	8	91	6.7%	7.9%	90.1%	70.0%
8	120	960	80.0%	6	97	5.0%	5.9%	96.0%	80.0%
9	120	1080	90.0%	3	100	2.5%	3.0%	99.0%	90.0%
10	120	1200	100.0%	1	101	0.8%	1.0%	100.0%	100.0%
	1200			101		8.4%			

Figure 2: Lift and cumulative lift for RFM and logistic regression

Refer to *Figure 3*.

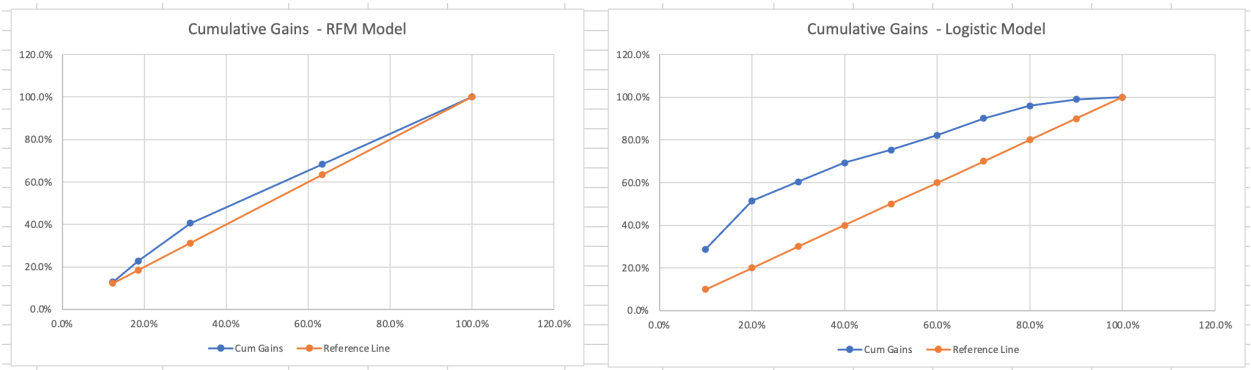

Figure 3: Lift and cumulative lift for RFM and logistic regression

Part IV: Targeting

Question 11

What is the breakeven response rate? (Hint: No R is required.)

The breakeven rate is \$1/\$10.2 = 9.8%

```
breakeven_rate = 1/10.2
```

Question 12

Using your logistic regression result, for the customers in the validation dataset, create a new variable (call it "target") with a value of 1 if the customer's predicted probability is greater than or equal to the breakeven response rate and 0 otherwise.

By comparing the predicted probability and the breakeven rate, 320 customers are selected as the target customers.

```
breakeven_rate = 1/10.2
data_val$target <- ifelse(data_val$probs >= breakeven_rate, 1, 0)
head(data_val$target,10)
```

```
##  [1] 0 1 0 1 0 0 0 1 0 0
```

```
sum(data_val$target,na.rm = TRUE)
```

```
## [1] 320
```

Question 13

Using your logistic regression result, compute the following numbers

```
# Compute percentage of customers to target
target_percentage <- sum(data_val$target == 1,na.rm = TRUE) / 1200
cat("Percentage of customers to target:", round(target_percentage * 100, 2), "%\n")
```

```
## Percentage of customers to target: 26.67 %
```

```
# Compute average response rate among targeted customers
response_rate <- mean(data_val$probs[data_val$target == 1],na.rm = TRUE)
cat("Average response rate among targeted customers:", round(response_rate * 100, 2), "%\n")
```

```
## Average response rate among targeted customers: 22.26 %
```

```
# Compute expected number of buyers
expected_buyers <- 20000*target_percentage*response_rate
cat("Expected number of buyers:", expected_buyers, "\n")
```

```
## Expected number of buyers: 1187.112
```

```
# Compute gross profit
gross_profit <- expected_buyers * 10.20
cat("Gross profit:", round(gross_profit, 2), "\n")
```

```
## Gross profit: 12108.55
```

```r
# Compute gross sales
gross_sales <- expected_buyers * (10.2+1)
cat("Gross sales:", round(gross_sales, 2), "\n")
```

```
## Gross sales: 13295.66
```

```r
# Compute total marketing cost
marketing_cost <- 20000*target_percentage * 1
cat("Total marketing cost:", round(marketing_cost, 2), "\n")
```

```
## Total marketing cost: 5333.33
```

```r
# Compute marketing ROI
marketing_roi <- gross_profit / marketing_cost
cat("Marketing ROI:", round(marketing_roi * 100, 2),"%\n")
```

```
## Marketing ROI: 227.04 %
```

Question 14

Using your R(FM) result, compute the following numbers

```r
# Find deciles with predicted response rate greater than breakeven rate
deciles_above_breakeven <- which(book_data$Avg_Response_Rate > breakeven_rate)
deciles_above_breakeven
```

```
## [1] 2 3
```

```r
deciles_of_interest <- which(book_data$Avg_Response_Rate > breakeven_rate)
cat("Deciles with predicted response rate greater than breakeven rate:", deciles_of_interest, "\n")
```

```
## Deciles with predicted response rate greater than breakeven rate: 2 3
```

```r
data_val$target_rfm <- ifelse(data_val$R_quintiles %in% deciles_of_interest, 1, 0)

# Compute percentage of customers to target
target_percentage_rfm <- sum(data_val$target_rfm == 1,na.rm = TRUE) / 1200
cat("Percentage of customers to target:", round(target_percentage_rfm * 100, 2), "%\n")
```

```
## Percentage of customers to target: 18.92 %
```

```r
# Compute average response rate among targeted customers
response_rate_rfm <- sum(data_val$Choice[data_val$target_rfm == 1],na.rm = TRUE)/sum(data_val$target_rfm
cat("Average response rate among targeted customers:", round(response_rate_rfm * 100, 2), "%\n")
```

```
## Average response rate among targeted customers: 12.33 %
```

```r
# Compute expected number of buyers
expected_buyers_rfm <- 20000*target_percentage_rfm*response_rate_rfm
cat("Expected number of buyers:", expected_buyers_rfm, "\n")
```

```
## Expected number of buyers: 466.6667
```

```r
# Compute gross profit
gross_profit_rfm <- expected_buyers_rfm * 10.20
cat("Gross profit:", round(gross_profit_rfm, 2), "\n")
```

```
## Gross profit: 4760
```

```r
# Compute gross sales
gross_sales_rfm <- expected_buyers_rfm * (10.2+1)
cat("Gross sales:", round(gross_sales_rfm, 2), "\n")
```

```
## Gross sales: 5226.67
```

```r
# Compute total marketing cost
marketing_cost_rfm <- 20000*target_percentage_rfm * 1
cat("Total marketing cost:", round(marketing_cost_rfm, 2), "\n")
```

```
## Total marketing cost: 3783.33
```

```r
# Compute marketing ROI
marketing_roi_rfm <- gross_profit_rfm / marketing_cost_rfm
cat("Marketing ROI:", round(marketing_roi_rfm * 100, 2),"%\n")
```

```
## Marketing ROI: 125.81 %
```

Question 15

Compare the results of mass-marketing, pure-R model and logistic regression in terms of Marketing ROI, based on the calculations above.

First of all, mass marketing statistics are calculated as below:

```r
# Compute average response rate among targeted customers
response_rate_mass <- sum(data_val$Choice[data_val$Choice == 1], na.rm = TRUE)/1200
cat("Average response rate among targeted customers:", round(response_rate_mass * 100, 2), "%\n")
```

```
## Average response rate among targeted customers: 8.42 %
```

```r
# Compute expected number of buyers
expected_buyers_mass <- 20000*response_rate_mass
cat("Expected number of buyers:", expected_buyers_mass, "\n")
```

```
## Expected number of buyers: 1683.333
```

```r
# Compute gross profit
gross_profit_mass <- expected_buyers_mass * 10.20
cat("Gross profit:", round(gross_profit_mass, 2), "\n")
```

```
## Gross profit: 17170
```

```r
# Compute gross sales
gross_sales_mass <- expected_buyers_mass * (10.2+1)
cat("Gross sales:", round(gross_sales_mass, 2), "\n")
```

```
## Gross sales: 18853.33
```

```r
# Compute total marketing cost
marketing_cost_mass <- 20000* 1
cat("Total marketing cost:", round(marketing_cost_mass, 2), "\n")
```

```
## Total marketing cost: 20000
```

```r
# Compute marketing ROI
marketing_roi_mass <- gross_profit_mass / marketing_cost_mass
cat("Marketing ROI:", round(marketing_roi_mass * 100, 2),"%\n")
```

```
## Marketing ROI: 85.85 %
```

It can be concluded that using logistic regression yields the largest marketing ROI (227.04 %), followed by the RMF model (125.81 %). The worst performing approach is mass marketing, only creating a marketing ROI of 85.85 %.