데이터베이스 마케팅

2nd 과제 (RFM-exercise)

비지니스 애널리틱스

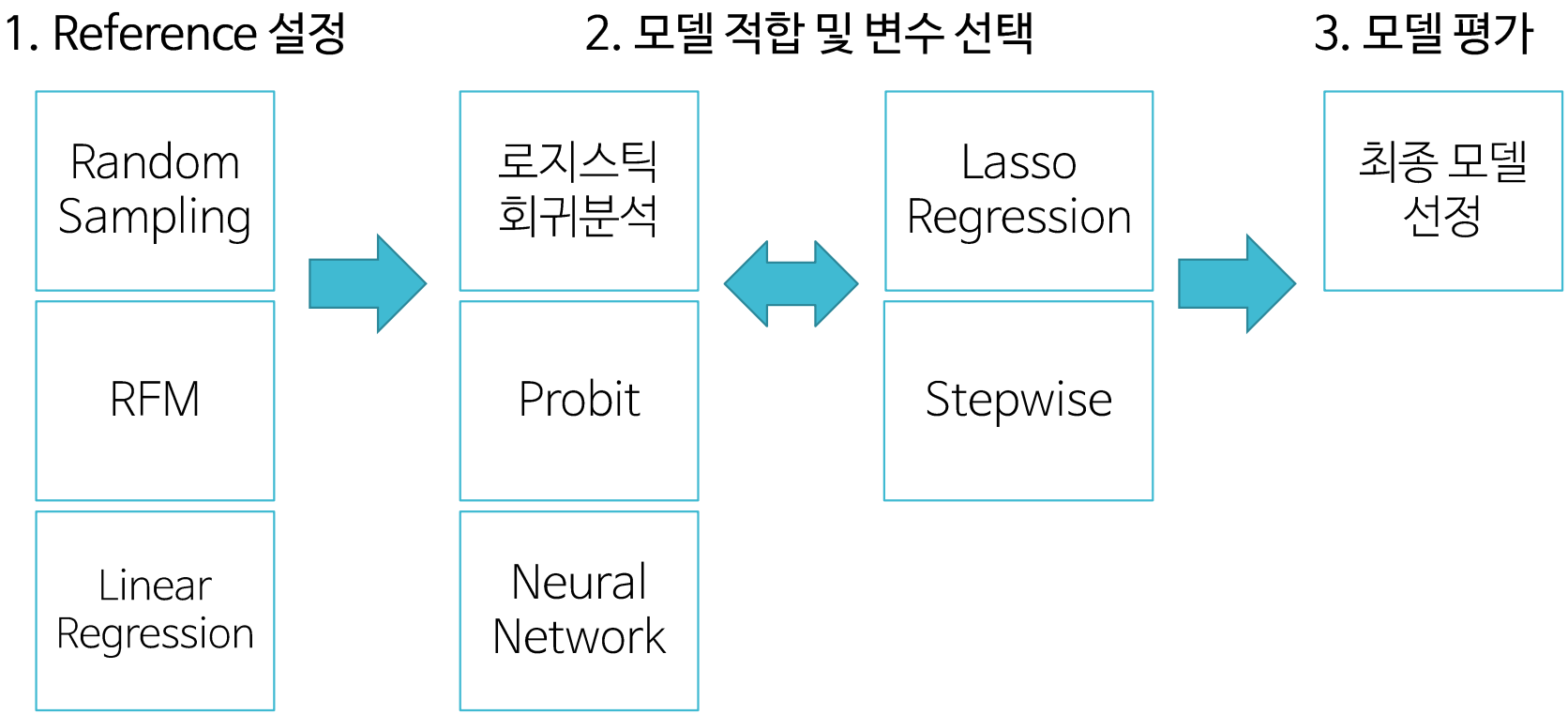
신 미 영, 장 재 석, 이 규 봉

목차

0. 개요

카탈로그 발송 대상 2000명(Validation Set)의 고객 중, 가장 구매 가능성이 높은 500명을 예측하는 모형을 도출해 보았습니다.

Random Sampling 및 RFM 모델로 Reference 예측률을 계산해, 이번 프로젝트의 최저기준으로 설정 하였습니다. 조원들이 로지스틱 회귀분석, Lasso Regression, Neural network, Probit 모형을 시도해 보았으며, Lasso 와 Stepwise 를 통해 변수 선택을 진행했습니다. 선택된 변수를 각각의 모델에 다시 적합하며, 최상의 예측률이 나오는 모델을 선정하였습니다 .



1. Reference Model

분석에 앞서 각 변수들의 특성을 파악합니다.

변수명	내용
Customer id	고객식별 **ID**
Gender (M = male; F = female)	성별 (**M** : 남성, **F** : 여성)
Monetary	총 구매 금액
Recency	최근 구매 후 기간 (개월)
Frequency	구매 빈도
Duration	첫 구매 후 기간 (개월)
Purchase code	구매 코드 (**1** : 구매; **2** : 비구매)

summarize monetary recency frequency duration

Variable	Obs	Mean	Std. Dev.	Min	Max
monetary	4,000	208.6625	100.7279	15	478
recency	4,000	13.368	8.161448	2	36
frequency	4,000	3.848	3.480578	1	12
duration	4,000	26.41075	18.27288	2	99

tabulate gender

gender	Freq.	Percent	Cum.
F	2,827	70.67	70.67
M	1,173	29.32	100.00
Total	4,000	100.00	

tabulate purchase

purchase	Freq.	Percent	Cum.
0	3,675	91.88	91.88
1	325	8.13	100.00
Total	4,000	100.00	

1. Reference Model

Q1. Randomly select 500 customers in the validation sample

Variable	Obs	Mean	Std. Dev.	Min	Max
purchase	500	.068	.251998	0	1

랜덤 샘플링으로 500명을 선정한 결과,
구매한 고객의 비율이 6.8% 입니다.

1. Reference Model

Q2. Create 2 x 2 x 2 RFM codes for estimation and validation sample.

$$R = \begin{cases} 1 \ if \ recency > 12 \\ 2 \ if \ recency \le 12 \end{cases} \quad F = \begin{cases} 1 \ if \ frequency < 3 \\ 2 \ if \ frequency \ge 3 \end{cases} \quad M = \begin{cases} 1 \ if \ monetary < 209 \\ 2 \ if \ monetary \ge 209 \end{cases}$$

위 조건대로 RFM을 적용하여 proportion 구하기

subpop_7: 2 2 1
subpop_8: 2 2 2

```
------------------------------------------------------------------------------
      Over | Proportion   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _subpop_7 |   .1751825   .0325953     .1200733    .2484434
 _subpop_8 |   .1672131   .0214025     .1293263    .2134777
```

Validation Set에 적용

subpop_7: 2 2 1- validation 자료 1-400까지
Subpop_8: 2 2 2 - validation 자료 401~699 중 100개 random sample로 추출

```
 Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   purchase |       400         .16    .3670652         0          1

 Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   purchase |       100          .1    .3015113         0          1
```

((400*0.16)+(100*0.1))/500=0.148

500명 중 74명이 구매해 14.8%의 예측률을 보입니다.

1. Reference Model

Q3. Create 5 x 5 x 5 RFM codes for estimation and validation sample.

$$R=\begin{cases}1\ if\ recency>16\\ 2\ if\ 12<recency\le 16\\ 3\ if\ 8<recency\le 12\\ 4\ if\ 4<recency\le 8\\ 5\ if\ recency\le 4\end{cases}\qquad F=\begin{cases}1\ if\ frequency=1\\ 2\ if\ frequency=2\\ 3\ if\ 2<frequency\le 5\\ 4\ if\ 5<frequency\le 9\\ 5\ if\ frequency>9\end{cases}\qquad M=\begin{cases}1\ if\ monetary\le 113\\ 2\ if\ 113<monetary\le 181\\ 3\ if\ 181<monetary\le 242\\ 4\ if\ 242<monetary\le 299\\ 5\ if\ monetary>299\end{cases}$$

위 조건대로 RFM을 적용하여 promotion 구하고 Validation Set에 적용

validation 자료 1-492

Variable	Obs	Mean	Std. Dev.	Min	Max
purchase	492	.1219512	.3275625	0	1

validation 자료 493~ 525중 8개 random sample로 추출

Variable	Obs	Mean	Std. Dev.	Min	Max
purchase	8	.25	.46291	0	1

((492*.1219512)+(8*0.25))/500=0.124

500명 중 62명이 구매해 12.4%의 예측률을 보입니다.

1. Reference Model

Q4. Linear regression

$$\text{Purchase} = \beta_0 + \beta_1(recency) + \beta_2(frequency) + \beta_3(monetary)$$

```
reg purchase recency frequency monetary
      Source |       SS           df       MS      Number of obs   =     2,000
-------------+----------------------------------   F(3, 1996)      =     22.82
       Model |  4.96472782         3  1.65490927   Prob > F        =    0.0000
    Residual |  144.750772     1,996  .072520427   R-squared       =    0.0332
-------------+----------------------------------   Adj R-squared   =    0.0317
       Total |  149.7155     1,999  .074895198   Root MSE        =    .2693

------------------------------------------------------------------------------
    purchase |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     recency |  -.0043872   .0007437    -5.90   0.000    -.0058458   -.0029287
   frequency |   .0087193   .0019768     4.41   0.000     .0048425    .0125962
    monetary |    .000065   .0000689     0.94   0.345      -.00007    .0002001
       _cons |   .0921429   .0170722     5.40   0.000     .0586617    .1256241
------------------------------------------------------------------------------

predict yhat
su purchase in 1/500

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    purchase |       500         .16    .3669732         0          1
```

$$\hat{y} = 0.09214 - 0.00438 * recency + 0.00872 * frequency + 0.00007 * monetary$$

500명 중 80명이 구매해 16%의 예측률을 보입니다.

2. 로지스틱 회귀분석(1)

Validation Set에서 Purchase가 1일 확률을 얻어 상위 500명을 선정하고자, 모든 변수를 활용한 회귀분석을 시행하였습니다.

(결과 예측에 영향이 없는 id 변수는 모든 과정에서 제외)

```
Call:
glm(formula = purchase ~ ., family = binomial, data = trset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0272  -0.4445  -0.3461  -0.2645   3.0157

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.1057486  0.3429606  -9.056  < 2e-16 ***
gender       0.6446095  0.1723430   3.740 0.000184 ***
monetary     0.0010490  0.0009594   1.093 0.274183
recency     -0.0673636  0.0169854  -3.966 7.31e-05 ***
frequency    0.1388373  0.0540685   2.568 0.010234 *
duration    -0.0098151  0.0111535  -0.880 0.378857
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1129.7  on 1999  degrees of freedom
Residual deviance: 1046.2  on 1994  degrees of freedom
AIC: 1058.2

Number of Fisher Scoring iterations: 6
```

2. 로지스틱 회귀분석(1)

모델 평가

Training Set에서 적합된 모델로 Validation Set의 500명을 선정해, purchase가 1인 고객 수 비율로 평가 합니다.

	gender	monetary	recency	frequency	duration	purchase	prob
3773	2	412	2	11	28	0	0.4337259
2699	2	341	2	12	40	1	0.4206533
2790	2	264	4	12	30	1	0.3923502
3067	2	408	2	11	46	0	0.3899471
3414	2	304	6	12	38	1	0.3523488
3901	2	282	2	11	50	0	0.3500148

```
FALSE  TRUE
  411    89
```

$$P(Y=1 \mid X_1, X_2, X_3, X_4, X_5) = \frac{\exp(-3.105 + 0.644 * G + 0.001 * M - 0.673 * R + 0.138 * F - 0.009 * D)}{1 + \exp(-3.105 + 0.644 * G + 0.001 * M - 0.673 * R + 0.138 * F - 0.009 * D)}$$

회귀분석으로 도출된 모델을 Validation Set에 적용해 500명을 선정한 결과, 89명의 구매로 17.8%의 예측률을 보였습니다.

3. Probit 분석(1)

예측율을 향상시키고자 다른 분석법을 적용해 보았습니다. 모든 변수를 활용해 Probit 분석을 수행합니다.

Probit 모형은 링크 함수가 'probit 함수'라는 점에서 logit 모형과 차이점이 있습니다.

```
Probit regression                               Number of obs     =      2,000
                                                LR chi2(5)        =     107.87
                                                Prob > chi2       =     0.0000
Log likelihood = -508.47659                     Pseudo R2         =     0.0959

------------------------------------------------------------------------------
    purchase |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |  -.5189626   .0876546    -5.92   0.000    -.6907624   -.3471627
     recency |  -.0427152   .0091945    -4.65   0.000    -.0607361   -.0246942
   frequency |   .0463073   .0300231     1.54   0.123    -.0125369    .1051515
    monetary |   .0007673   .0005021     1.53   0.126    -.0002169    .0017514
    duration |   .0010854   .0062973     0.17   0.863    -.0112571     .013428
       _cons |   -.966089   .1359776    -7.10   0.000     -1.2326   -.6995777
------------------------------------------------------------------------------
```

```
. predict pprobit1
. su purchase in 1/500

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    purchase |       500        .186    .3894964         0          1
```

$$Pr(Y=1|X_1, X_2, X_3, X_4, X_5) = \Phi(-0.966089 - 0.5189626 * gender - 0.0427152 * recency + 0.0463073 * frequency + 0.0007673 * monetary + 0.0010854 * duration)$$

Probit 분석으로 500명의 고객 중 93명이 구매, 18.6%의 예측률을 보였습니다.

4. Neural network

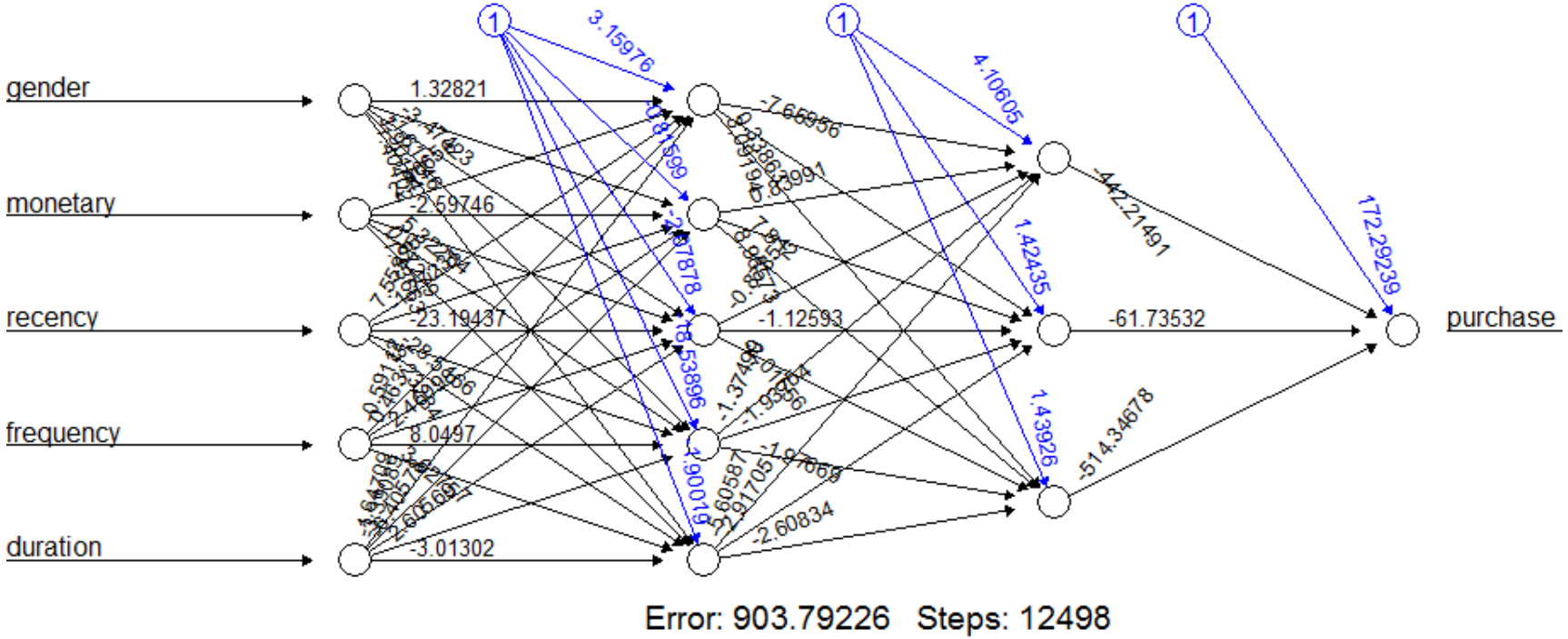
또 다른 분석법으로 Neural Network 모형을 시도해보았습니다.

```
library(neuralnet)
nn <- neuralnet(purchase ~ gender + monetary
+ recency + frequency + duration,
data=scaled_training_data_df,hidden=c(5,3),
linear.output=FALSE)
```

```
Summary(nn)
Length Class     Mode
call              5 -none-    call
response       2000 -none-    numeric
covariate     10000 -none-    numeric
model.list        2 -none-    list
err.fct           1 -none-    function
act.fct           1 -none-    function
linear.output     1 -none-    logical
data              7 data.frame list
net.result        1 -none-    list
weights           1 -none-    list
startweights      1 -none-    list
generalized.weights   1 -none-    list
result.matrix    55 -none-    numeric
```

4. Neural network

Neural network로 모델을 적합해 보았으나 의미있는 변수인지의 판단이 쉽지 않았습니다.
또한 최적의 Hidden Layer 를 갖춘 모형계산에 Computing Power가 많이 필요하다는 문제를 해결하지 못해, 결과 값을 확실을 갖고 참고할 수 없었습니다.



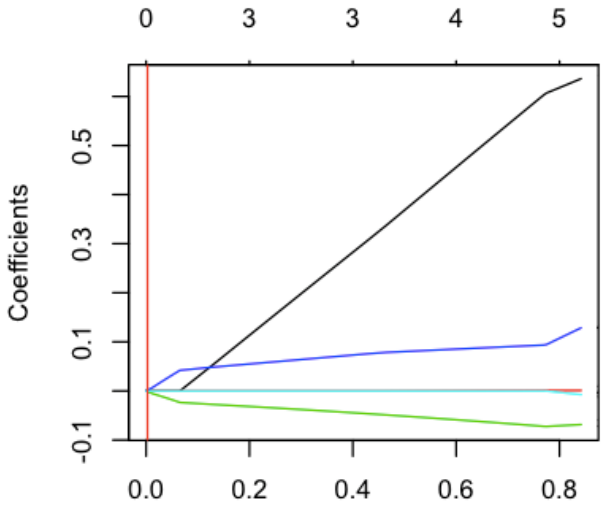
〈Neural network Model〉

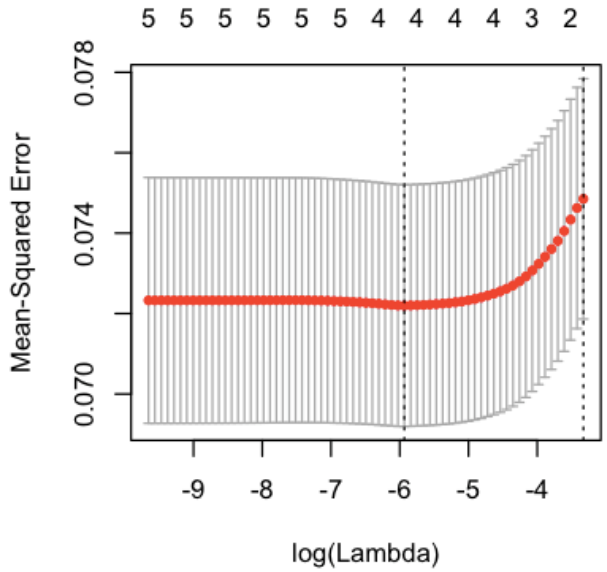
Hidden layer	3,3	5,2	***5,3***	5,4
예측율	15.2 %	13.4 %	***18 %***	12.8%

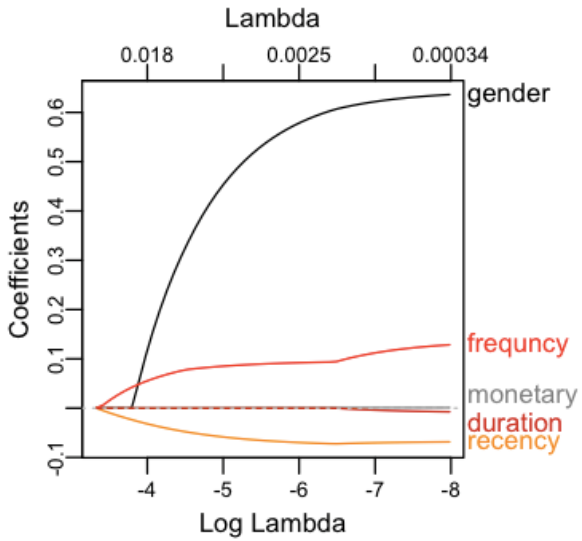
〈Hedden Layer 에 따른 예측률 〉

5. Lasso

‘모든 변수를 활용한 Probit 분석’의 예측률 18.6% 보다 개선을 이루고자, 변수 선택 과정을 기획했습니다. Lasso Regression으로 변수 선택 및 계수 추정을 함께 수행합니다.







MSE를 최소화하는 Lambda 값을 탐색해, 해당 지점의 변수별 계수를 추정합니다.

5. Lasso

모델 평가

Training Set에서 적합된 모델로 Validation Set의 500명을 선정해, purchase가 1인 고객 수 비율로 평가 합니다.

```
6 x 1 sparse Matrix of class "dgCMatrix"
                       1
(Intercept) -2.9723436609
gender       0.5734530655
monetary     0.0008154497
recency     -0.0693578550
frequency    0.0918079243
duration     .
```

$$\hat{y} = -2.972 + 0.573 * gender + 0.008 * monetary - 0.069 * recency + 0.091 * frequency$$

	gender	monetary	recency	frequency	duration	purchase	prob
2699	2	341	2	12	40	1	0.4206533
3773	2	412	2	11	28	0	0.4337259
3067	2	408	2	11	46	0	0.3899471
2221	2	253	2	12	70	0	0.3302925
2102	2	401	4	12	70	0	0.3348497
3901	2	282	2	11	50	0	0.3500148

```
FALSE  TRUE
  408    92
```

Lasso Regression 결과 duration 변수가 탈락됨을 알 수 있습니다.
적합된 모델을 Validation Set에 적용해 500명을 선정한 결과,
92명이 구매해 18.4%의 예측률을 보였습니다.

6. 로지스틱 회귀분석(2)

Lasso 수행 결과,
유의한 변수로 선택되지 못한 duration을 제외 후 로지스틱 회귀분석을 수행합니다.

```
Call:
glm(formula = purchase ~ ., family = binomial, data = trset4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9653  -0.4458  -0.3470  -0.2642   3.0250

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.0897186  0.3429373  -9.010  < 2e-16 ***
gender       0.6520261  0.1720912   3.789 0.000151 ***
monetary     0.0010826  0.0009584   1.130 0.258655
recency     -0.0768863  0.0131460  -5.849 4.96e-09 ***
frequency    0.0962040  0.0251020   3.833 0.000127 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1129.7  on 1999  degrees of freedom
Residual deviance: 1047.0  on 1995  degrees of freedom
AIC: 1057

Number of Fisher Scoring iterations: 6
```

6. 로지스틱 회귀분석(2)

모델 평가

Training Set에서 적합된 모델로 Validation Set의 500명을 선정해, purchase가 1인 고객 수 비율로 평가 합니다.

	gender	monetary	recency	frequency	purchase	prob
2699	2	341	2	12	1	0.3975163
3773	2	412	2	11	0	0.3928938
3067	2	408	2	11	0	0.3918613
2102	2	401	4	12	0	0.3764505
2221	2	253	2	12	0	0.3749372
3901	2	282	2	11	0	0.3598761

```
FALSE  TRUE
  408    92
```

$$P(Y=1 \mid X_1, X_2, X_3, X_4) = \frac{\exp(-3.089 + 0.652 * G + 0.001 * M - 0.076 * R + 0.096 * F)}{1 + \exp(-3.089 + 0.652 * G + 0.001 * M - 0.076 * R + 0.096 * F)}$$

Duration 변수를 제외 후 회귀분석으로 적합된 모델입니다.
Validation Set에 적용해 500명을 선정한 결과, 92명의 구매로 18.4%의 예측률을 보였습니다.

7. 로지스틱 회귀분석(3)

다른 방법의 변수 선택을 수행하고자 stepwise 기법을 이용합니다. 선정된 변수들로 로지스틱 회귀분석을 수행합니다.

```
Step:  AIC=1056.24
purchase ~ gender + recency + frequency

            Df Deviance    AIC
<none>           1048.2 1056.2
+ monetary   1   1047.0 1057.0
+ duration   1   1047.4 1057.4
- gender     1   1061.6 1067.6
- frequency  1   1075.3 1081.3
- recency    1   1089.9 1095.9
```

Stepwise 기법 수행 결과 *gender*, *frequency*, *recency* 가 유의미한 변수로 선정되었습니다.

7. 로지스틱 회귀분석(3)

다른 방법의 변수 선택을 수행하고자 stepwise 기법을 이용합니다.
선정된 변수들로 로지스틱 회귀분석을 수행합니다.

```
Call:
glm(formula = purchase ~ ., family = binomial, data = trset5)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9994  -0.4412  -0.3492  -0.2651   3.0044

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.90285    0.29855  -9.723  < 2e-16 ***
gender       0.64005    0.17166   3.729 0.000193 ***
recency     -0.07697    0.01315  -5.853 4.82e-09 ***
frequency    0.11187    0.02098   5.332 9.71e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1129.7  on 1999  degrees of freedom
Residual deviance: 1048.2  on 1996  degrees of freedom
AIC: 1056.2

Number of Fisher Scoring iterations: 6
```

7. 로지스틱 회귀분석(3)

모델 평가

Training Set에서 적합된 모델로 Validation Set의 500명을 선정해, purchase가 1인 고객 수 비율로 평가 합니다.

	gender	recency	frequency	purchase	prob
2221	2	2	12	0	0.3931214
2699	2	2	12	1	0.3931214
3067	2	2	11	0	0.3667735
3773	2	2	11	0	0.3667735
3901	2	2	11	0	0.3667735
2102	2	4	12	0	0.3570571

```
FALSE  TRUE
  409    91
```

$$P(Y=1 \mid X_1, X_2, X_3) = \frac{\exp(-2.902 + 0.640 * G - 0.076 * R + 0.111 * F)}{1 + \exp(-2.902 + 0.640 * G - 0.076 * R + 0.111 * F)}$$

Gender, recency, frequency 변수로 이루어진 회귀분석으로 도출된 모델입니다. Validation Set에 적용해 500명을 선정한 결과, 91명의 구매로 18.2%의 예측률을 보였습니다.

8. Probit 분석(2)

Stepwise 절차로 선정된 변수 gender, recency, frequency를 활용해 Probit 분석을 수행합니다.

```
Probit regression                               Number of obs     =      2,000
                                                LR chi2(3)        =     105.46
                                                Prob > chi2       =     0.0000
Log likelihood = -509.68032                     Pseudo R2         =     0.0938

------------------------------------------------------------------------------
    purchase |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   -.5181301   .0875203    -5.92   0.000    -.6896668   -.3465935
     recency |   -.0411875   .0064246    -6.41   0.000    -.0537795   -.0285955
   frequency |    .0617611   .0114939     5.37   0.000     .0392334    .0842887
       _cons |   -.8535164   .1077908    -7.92   0.000    -1.064782   -.6422504
------------------------------------------------------------------------------
```

```
predict pprobit4
. su purchase in 1/500

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    purchase |        500        .192    .3942675         0          1
```

$Pr(Y=1|X_1, X_2, X_3) =$
$\Phi(-0.8535164 - 0.5181301 * gender - 0.0411875 * recency + 0.0617611 * frequency)$

Probit 분석으로 500명의 고객 중 96명이 구매, 19.2%의 가장 높은 예측률을 보였습니다.

9. Conclusion

목차	선택 모델	선택 변수	예측률
1	Random Sampling	-	6.8 %
1	RFM Linear Regression	recency, frequency, monetary	16 %
2	Logistic Regression	gender, monetary, recency, frequency, duration	17.8 %
3	Probit Model	gender, monetary, recency, frequency, duration	18.6 %
4	Neural Network	gender, monetary, recency, frequency, duration	18 %
5	Lasso Regression	gender, monetary, recency, frequency	18.4 %
6	Logistic Regression	gender, monetary, recency, frequency	18.4 %
7	Logistic Regression	gender, recency, frequency	18.2 %
8	**Probit Model**	**gender, recency, frequency**	**19.2 %**

9. Conclusion

본 분석은 카달로그 보낼 2000명의 고객 중 500명을 선정하기 위해 데이터의 탐색적 분석을 실시하고 예측률이 가장 높은 모델을 찾는 것을 목적으로 하고 있습니다.

랜덤 샘플 , RFM 모델 그리고 선형회귀분석으로 Reference 예측률을 계산해, 이번 프로젝트의 최저기준으로 설정 하였습니다. 로지스틱 회귀분석, Lasso Regression, Neural network, Probit 모형을 시도해 보았으며, Lasso 와 Stepwise 를 통해 변수 선택을 진행한 결과 gender, recency, frequency 변수로 Probit 모델을 적합한 케이스가 19.2%의 가장 높은 구매 예측률을 기록했습니다.

위 자료와 같이 종속변수가 0,1인 이항분포 형태인 경우 대표적으로 로지스틱 회귀분석과 Probit 분석을 많이 사용하고 있습니다. 둘 중 어느 것이 더 좋다 말하기 어렵기 때문에 주워진 자료에 따라 모든 모형을 만들어서 예측률이 높거나 테스트 에러가 낮은 모형을 선택하는 것이 바람 직 할 것입니다.