R 실습 - 사례

과제 풀이

18.3 테이코 소프트웨어 카탈로그 판매회사

테이코 소프트웨어 카탈로그 판매회사

구매가능성이 높은 대상자들을 선택하여 카탈로그를 발송하기 위해 사용할 모델을 개발하려고 한다.

1. 각 카탈로그는 발송까지 약 2달러가 소요된다(인쇄, 우표 그리고 발송 비용 포함). 고객 데이터베이스에서 무작위로 선택한다면 나머지 180,000명에게 회사가 기대할 수 있는 총 수익을 추정해 보시오

- $103 \times (180{,}000 \times 0.053) - 2 \times 180{,}000 = 622{,}620$

  고객 1명당 평균 구매 금액 × (180,000 명 중 평균 구매자의 수) – 180,000명에게 발송하는 비용

2. 구매자나 혹은 비구매자로 고객을 분류하여 모형을 만드시오

- *각 케이스마다 t값 800개, v값 700개, s개 500개를 할당한 후 분할 변수에 따라 데이터를 학습자료로 추출하시오.*
- *완전한 로지스틱 회귀 모형을 이행하고, 변수의 최량 부분 집합을 선택하고, 그 변수를 이용하여 데이터에서 구매자와 비구매자를 찾는 회귀모형을 실행하시오(로지스틱 회귀는 예상 '구매확률 ' 을 산출해주기 때문에 사용하는데 이것은 나중에 분석 작업에서 필요하다).*

- 변수의 최량 부분 집합은 freq, source_h1, web.order1, source_c1, source_u1, address_is_res1, source_b1, source_a1, source_w1을 포함

  $\text{logit}(Y) = -3.3761 + 2.5660 \times freq - 4.2013 \times source_{h1} + 0.7987 \times web.order1 - 1.0444 \times source_{c1} + 1.1590 \times source_{u1} - 0.7443 \times address\ is\ res1 - 0.8460 \times source_{b1} + 0.8149 \times source_{a1} + 0.6137 \times source_{w1}$

테이코 소프트웨어 카탈로그 판매회사

구매가능성이 높은 대상자들을 선택하여 카탈로그를 발송하기 위해 사용할 모델을 개발하려고 한다

3. 구매자들의 지출을 예측할 수 있는 모형을 개발하시오.

- *데이터 시트의 복사본(데이터2)을 만들어서 Purchase 변수에 따라 정렬하고 Purchase 값이 0인 기록들을 제거한다(결과적으로 데이터2에는 구매자만 남는다)*
- *분할 변수에 근거해서 데이터세트를 학습용과 검증용으로 나누시오.*
- *두 가지 기법(다중 선형회귀 / 회귀나무 모형)을 사용하여 지출 예측 모형을 개발하시오.*
- *검증 데이터 평가 결과를 기준으로 하나의 모형을 선택하시오.*

▶ 검증 데이터로 각 모델에 rmse를 계산한 결과 값이 더 작은 다중 선형회귀 선택(다중 선형회귀: 163.281, 회귀나무: 185.771)

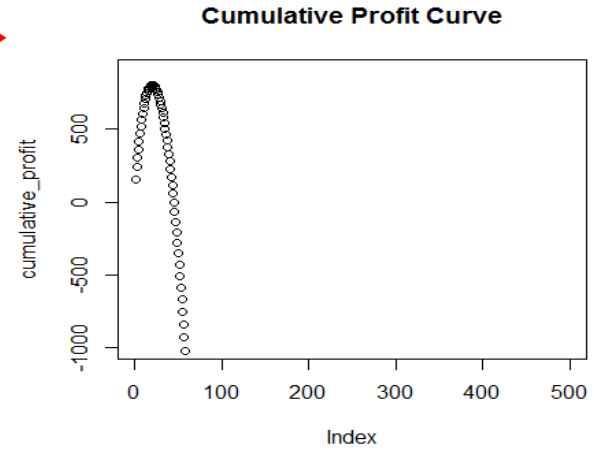
다중 선형회귀 모델:

$$Y = 40.089 + 45.805 \times US1 + 51.116 \times source_{r1} + 92.325 \times Freq - 0.027 \times last\ update\ days\ ago - 94.043 \times address\ is\ res1 - 87.059 \times source_{m1}$$

테이코 소프트웨어 카탈로그 판매회사

4. 원본 평가 데이터 분할로 돌아가시오. 이 평가용 데이터는 구매자와 비구매자를 모두 포함하고 있다는 것을 유념하시오. 비록 이 평가용 데이터가 선택된 분류 모형의 평가 점수를 포함하고 있지만 아직까지 이 분할을 사용하지 않았으므로 우리 모형들의 수행 평가로부터 공정한 값을 기대할 수 있다. 작업하는 시트의 평가용 데이터 부분의 카피를 만드는 것이 최선의 방법인데 그것에 분석을 적용할 것이기 때문이다. 이 카피는 스코어 분석이라고 불린다.

- *이 시트에 평가용 데이터 분류에서 나온 '성공 예상 확률'(성공=Purchase) 칼럼을 복사하시오.*
- *이 시트에 선택된 예측 모델의 점수를 기록하시오.*
- *구매 예상 확률(성공) / 실제 지출(달러) / 예상 지출(달러) 칼럼들을 서로 인접할 수 있도록 배치하시오.*
- *'구매 예상 확률 ' 에 0.107을 곱한 '수정 구매 확률' 칼럼을 추가하시오. 구매자를 과다표집 않도록 하기 위함이다.*
- *기대 지출 칼럼을 추가하시오(수정 구매 확률 X 예상 지출).*
- *'예상 지출' 칼럼을 기준으로 모든 기록들을 정렬하시오.*
- *누적 향상도를 계산하시오(=누적 '실제 지출 ' 을 평균 지출로 나눈다).*
- *이 누적 향상 곡선을 사용하여 데이터 마이닝 모형을 근거로 우송한 결과 얻어지는 최대 순이익을 추정하시오.*



그래프의 최대점(즉, 최대 순이익)은 805.791이다

테이코 소프트웨어 카탈로그 판매회사 - 풀이

- 데이터 로드 및 binary 변수 설정

```
library(xlsx)
library(data.table)

df <- read.xlsx('D:/DM_TA/Tayko.xls', 2)
colnames(df)
binary_col <- c(2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,21,22,23,24)
for (i in binary_col){
  df[[i]] <- factor(df[[i]])
}
```

- 데이터 분할을 통해 학습, 검증, 평가 데이터 생성

```
list_of_datset = split(df, df$Partition)
nrow(list_of_datset[[1]])
nrow(list_of_datset[[2]])
nrow(list_of_datset[[3]])
training_data <- list_of_datset[[2]]
validation_data <- list_of_datset[[3]]
test_data <- list_of_datset[[1]]
```

테이코 소프트웨어 카탈로그 판매회사 - 풀이

- 변수의 최량 부분 집합을 고려한 로지스틱 회귀 모형 생성

```
fullmod <- glm(Purchase ~ 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+last_update_days_ago+X1st_update_days_ago+
                 Web.order+Gender.male+Address_is_res, data = training_data, family = binomial(link=logit))
nothing <- glm(Purchase ~ 1, data = training_data, family = binomial(link=logit))
bothways <- step(nothing, list(lower=formula(nothing),upper=formula(fullmod)),direction="both",trace=0)
summary(bothways)
prediction <- predict(bothways, newdata = test_data)
logistic_result <- (10^prediction)/(1+10^prediction)
show(logistic_result)

Call:
glm(formula = Purchase ~ Freq + source_h + Web.order + source_c +
    source_u + Address_is_res + source_b + source_a + source_w,
    family = binomial(link = logit), data = training_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4924  -0.5393  -0.1707   0.5645   2.2303

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.3761     0.2902 -11.633  < 2e-16 ***
Freq              2.5660     0.2212  11.598  < 2e-16 ***
source_h1        -4.2013     0.6055  -6.939 3.95e-12 ***
Web.order1        0.7987     0.2026   3.943 8.06e-05 ***
source_c1        -1.0444     0.4287  -2.436  0.01485 *
source_u1         1.1590     0.3286   3.527  0.00042 ***
Address_is_res1  -0.7443     0.2829  -2.631  0.00852 **
source_b1        -0.8460     0.5304  -1.595  0.11071
source_a1         0.8149     0.3388   2.406  0.01614 *
source_w1         0.6137     0.2866   2.142  0.03222 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1106.83  on 799  degrees of freedom
Residual deviance:  612.12  on 790  degrees of freedom
AIC: 632.12

Number of Fisher Scoring iterations: 7
```

테이코 소프트웨어 카탈로그 판매회사 - 풀이

- Purchase 값이 0인 기록들을 제거 및 학습용, 검증용 데이터 생성

```
new_df <- df[!(df$Purchase==0),]
new_list_of_datset = split(new_df, new_df$Partition)
training_data <- new_list_of_datset[[2]]
validation_data <- new_list_of_datset[[3]]
setDT(training_data)[,c("sequence_number", "Purchase", "Partition", "NA.","NA..1"):=NULL]
setDT(validation_data)[,c("sequence_number","Purchase", "Partition", "NA.","NA..1"):=NULL]
```

- 지출 예측 모형(다중 선형 회귀 / 회귀나무 모형)

```
library(MASS)
fullmod <- lm(Spending ~ ., data = training_data)
summary(fullmod)
step <- stepAIC(fullmod, direction="both")
summary(step)
```

```
library(rpart)
regression_tree <- rpart(Spending ~ 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+last_update_days_ago+X1st_update_days_ago+
                           web.order+Gender.male+Address_is_res,method="anova",data=training_data)
plot(regression_tree); text(regression_tree)
p_regression_tree <- prune(regression_tree, cp = regression_tree$cptable[which.min(regression_tree$cptable[,"xerror"]),"CP"])
plot(p_regression_tree); text(p_regression_tree)
```

테이코 소프트웨어 카탈로그 판매회사 - 풀이

- 지출 예측 모형 평가

```
install.packages('Metrics', dependencies = TRUE)
library(Metrics)
actual_spending <- validation_data$Spending
prediction_ml <- predict(step, newdata = validation_data)
prediction_rt <- predict(p_regression_tree, newdata = validation_data)
rmse(prediction_ml, actual_spending)
rmse(prediction_rt, actual_spending)
```

```
Call:
lm(formula = Spending ~ US + source_r + Freq + last_update_days_ago +
    Address_is_res + source_m, data = training_data)

Residuals:
    Min      1Q  Median      3Q     Max
-441.51  -97.94  -19.33   68.56 1106.77

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          40.088876  29.376596   1.365  0.17319
US1                  45.805429  22.725345   2.016  0.04456 *
source_r1            51.115576  33.338519   1.533  0.12607
Freq                 92.325315   5.499509  16.788  < 2e-16 ***
last_update_days_ago -0.025709   0.008255  -3.115  0.00199 **
Address_is_res1      -94.043053  20.838332  -4.513 8.59e-06 ***
source_m1            -87.059305  59.673852  -1.459  0.14543
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 166.5 on 372 degrees of freedom
Multiple R-squared:  0.5102,    Adjusted R-squared:  0.5023
F-statistic: 64.59 on 6 and 372 DF,  p-value: < 2.2e-16
```

테이코 소프트웨어 카탈로그 판매회사 - 풀이

- 선택된 예측 모델의 점수, 수정 구매 확률, 기대 지출 칼럼 계산 및 추가

```
prediction_ml_test <- predict(step, newdata = test_data)
score_analysis <- setDT(test_data)[,c("sequence_number", "Partition", "NA.","NA..1"):=NULL]
logistic_result <- 0.107*logistic_result
score_analysis <- cbind(score_analysis, prediction_ml_test, logistic_result)
expected_spending <- prediction_ml_test * logistic_result
score_analysis <- cbind(score_analysis, expected_spending)
```

- 예상 지출' 칼럼을 기준으로 기록 정렬 및 누적 향상도 계산

```
score_analysis <- score_analysis[order(-expected_spending),]
cumulative_actual_spending <- sum(score_analysis$Spending)
average_spending <- sum(logistic_result * score_analysis$Spending)
cumulative_lift <- cumulative_actual_spending/average_spending
show(cumulative_lift)
```

테이코 소프트웨어 카탈로그 판매회사 - 풀이

- 누적 향상 곡선 그리기 및 최대 순이익 추정

```
present_profit <- 0
cumulative_profit <- vector()
show(score_analysis$expected_spending)
for(i in 1:length(score_analysis$expected_spending)){
  present_profit <- present_profit + (score_analysis$expected_spending[i]) - 2*i
  cumulative_profit[i] <- present_profit
}
plot(cumulative_profit, main = 'Cumulative Profit Curve', ylim=c(-25000,2000))
max(cumulative_profit)
```